Parametric and Semiparametric Binary Choice Estimators - Evidence from Monte Carlo Studies

Süß, Philipp

3 December 2012

Parametric and Semiparametric Binary Choice Estimators - Evidence from Monte Carlo Studies

December 3, 2012

Thesis to obtain the degree: "Master of Science in Quantitative Economics"
Student: Philipp Süß
Institution: Goethe University Frankfurt
Supervisor: Prof. Dr. Florian Heiss
Institution: Johannes Gutenberg University Mainz, Chair of Statistics and Econometrics

Abstract

The following thesis compares the performance of several parametric and semiparametric estimators in binary choice models using the method of Monte Carlo studies. Particularly, the thesis compares estimators of the parametric linear probability-, logit- and probit model, a model derived from Cauchy distributed errors (cauchit model) as well as the estimator proposed by Klein and Spady (KS) and the local likelihood logit estimator by Frölich (LLL), which are of the semiparametric class. Furthermore, the thesis proposes a Hausman type test to compare parametric with semiparametric estimators. The main results are as follows: all considered estimators delivered decent estimates of the average marginal effects, independent of the assumed functional form. The results for the estimation of marginal effects at specific points are different. The parametric estimators generally perform poorly, whereas the estimators derived from the true models perform well. Klein and Spady´s estimator performs decently in large samples. Moreover, a good performance with respect to the root mean squared error (RMSE) does generally not translate into a good estimation of the marginal effects.

Contents

List of Tables

List of Figures

1 Introduction

The following thesis compares the performance of several parametric and semiparametric estimators used in binary choice models. The aim is to give practical guidance for the use of estimators in applied econometric work. The main part of the thesis consists of a comparison of the different estimators via Monte Carlo studies. Additionally, a Hausman test is proposed as a guide to choose between several competing models.

The thesis is structured as follows: In the first chapter, I will give a brief introduction to the framework of binary choice models as well as a description of the quantities of interest. The first part of chapter two reviews the well known parametric estimators which are the ones derived from the linear probabilty-, the probit- and the logit model. Furthermore, a model derived from Cauchy distributed errors, from now on called "cauchit model", is introduced. The first part of chapter two ends with a discussion of how to construct ideal models for known distributions of the error term. The second part of chapter two is concerned with the class of semiparametric estimators. As will be discussed in greater detail in section 1.1, the choice of the functional form of the relationship between the dependent and independent variables often appears to be arbitrary. The advantage of the semiparametric estimators stems from the fact that they do not require assumptions on the functional form and therefore have the appeal to be more robust. This robustness comes in general at the cost of less efficiency. Since the theory of semi- and nonparametric estimation is less common than its parametric counterpart, the second part of chapter two starts with an introduction to semiparametric estimation, which is strongly related to the discussion in Cameron and Trivedi (2005, pp. 294-333). After introducing the semiparametric methodology, the estimator of Klein and Spady (1993) and the local likelihood logit estimator by Frölich (2006), as representatives of the class of semiparametric estimators, are described. Since the class of semiparametric estimators is large[1] the choice made deserves some justification. The estimator of Klein and Spady is chosen on theoretical grounds. It attains the asymptotic semiparametric efficiency bound and therefore seems to be an obvious choice for an estimator if the sample size is large. The choice of the local likelihood logit estimator is motivated by the result obtained by Frölich (2006) that his estimator outperforms Klein and Spady´s in several specifications. Chapter three describes the different setups of the Monte Carlo study. In total, I present the results of ten different Monte Carlo setups. These Monte Carlo setups can be divided into two groups. The first seven setups vary the distribution of the error term. In these setups, the link function is unknown and potentially misspecified, while the index is well specified. The errors are drawn from a normal-, logistic-, Cauchy-, Gumbel- and mixture normal distribution. Normal and logistic error terms are frequently assumed in applied work, whereas Cauchy distributed errors represent errors with fat tails. Gumbel distributed errors are skewed and errors from the mixture normal distribution are bimodal. Furthermore, one setup extends the number of non-constant regressors from three to six and one uses error terms which contain "outliers". The next three setups focus on misspecifications of the index given normally distributed errors. The

[1] An extensive list of semiparametric estimators is given in Pagan and Ullah (1999, pp. 272-299).

misspecifications result from omitted variables or misspecification of the index function. The results are presented in chapter four. I chose the following quantities to assess the performance of the estimators. The predictive power is mainly assessed by the root mean squared error, however additional measures are considered. The performance concerning the marginal effects is measured via the comparison of the true marginal effects and the estimated marginal effects. Here the average marginal effect, the marginal effect at the average and marginal effects at the first and third quartile are considered. After assessing the performance of the estimators chapter five proposes a Hausman type test as a potential decision rule between parametric and semiparametric estimators. Finally chapter six concludes.

1.1 Binary choice models

This class of models has the special feature that the dependent variable takes only two values. Such models are frequently used to study choice phenomena, e.g. why people smoke, why they are homeowners and not tenants or what drives people to become criminal, to name just a few. As usual in econometric analysis, the focus of the researcher lies either in a causal analysis or in predicting outcomes given certain characteristics. Since the conditional expectation coincides with conditional probabilities in binary choice models, it seems convenient that causal or predictive statements are based on probabilities.[2] An exemplary statement of an empirical analysis in a binary context is the following: increasing yearly income by 1000 € and holding all other characteristics of an individual constant increases the likelihood of being a house owner by 2.5 %. Since probabilites are bound between zero and one, the framework to study such phenomena should take this into account. While in principle these questions can be adressed by the linear regression framework, the fact that predicted probabilities may lie outside of the closed zero-one intervall for some observations makes the use of this econometric model problematic. In the following, I will give a brief technical introduction to the binary choice framework, which is closely related to the discussion in Cameron and Trivedi (2005, pp. 463-487).

The starting point of the empirical analysis is a random sample of independent identically distributed (i.i.d.) observations $\{y_i, X_i\}_{i=1}^N$ where y_i is a scalar dependent variable and X_i is a $k \times 1$ vector of explanatory variables. Since the observations are assumed to be i.i.d., data fitting the model presented is likely to be cross sectional. In the following, I will sometimes rely on matrix notation. Capital letters either denote vectors or matrices and lowercased letters scalars. Y is the $N \times 1$ vector of endogenous variables and X is the $N \times k$ matrix of regressors. As pointed out before, y_i can only take two values which are labelled as 0 and 1. Using the homeowner example, y_i could take the value 0 if the observed individual i is a tenant and 1 if he or she is a homeowner. Since y_i is a binary random variable its distribution is necessarily Bernoulli. The density function of a Bernoulli distribution is given by

[2] The conditional probabilities will be modelled as $p(X_i) = G(X_i'\beta)$, where $X_i'\beta = \sum_{j=1}^k x_{ij}\beta_j$ is called the index function. The function G is called the link function.

$$f(y_i) = \begin{cases} p_i \text{ for } y_i = 1 \\ 1 - p_i \text{ for } y_i = 0 \end{cases} \tag{1}$$

Due to the independence of the observations in the random sample the joint likelihood function of all N dependent variables is the product of the individual likelihood functions. The fact that y_i is assumed to have an identical distribution for all i implies that p is the same for all individuals ($p_i = p, \forall i$). Hence the joint likelihood becomes

$$\Psi(p) = \prod_{i=1}^{N} p^{y_i}(1-p)^{1-y_i} \tag{2}$$

The joint log-likelihood function is given by

$$L(p) = \sum_{i=1}^{N} [y_i ln(p) + (1 - y_i) ln(1-p)] \tag{3}$$

As common in econometric modelling, the researcher is concerned with modelling the conditional expectation of a random variable $E(Y|X)$ and how this conditional expectation changes when the explanatory variables change $\frac{dE(Y|X)}{dX_j}$. In binary choice models, the modelling of the conditional expectation is equivalent to modelling the conditional probabilities. This can be seen by noting that $E(Y|X) = 1 \cdot p(X) + 0 \cdot [1 - p(X)] = p(X)$.
A general econometric model is given by

$$E(Y|X) \quad = \quad p(X) = m(X) \tag{4}$$

where $m(X)$ is an arbitrary function. This model is nonparametric since it does not involve any parametrization (alternatively one can see the function as a collection of infinite parameters). Due to the fact that this thesis is concerned with parametric and semiparametric models the econometric model will look as follows:

$$E(Y|X) = p(X) = G(X'\beta_x) \tag{5}$$

where the function G is called link function and is assumed to be known in the parametric case. In the semiparametric case G is unknown and will be estimated simultaneously with the parameter vector β_x, which is of the dimension $k \times 1$. $X'\beta_x$ is called the index function. For all estimators considered in this work, with exception of the LLL estimator, the parameter vector is assumed to be constant across the whole sample. Formally stated, $\beta_x = \beta \ \forall x$.

The main advantage of semiparametric to nonparametric modelling is the reduction of the dimensionality of the link function. In the nonparametric case m is k-dimensional whereas G in equation (5) is a function of one variable.[3]

[3]Due to the high dimensionality, nonparametric estimators frequently cannot be estimated accurately in

Since the link function maps the regressors into the space of probabilities, it seems reasonable that the $image(G)$ does not contain elements outside $[0, 1]$ as it is required from Kolmogorovs axioms. Further desirable properties are smoothness and monotonicity. Smoothness easens the optimization required to obtain maximum likelihood estimates and monotonicity helps in interpreting the estimated coefficients, i.e. linking the sign of the coefficient on the parameter β_j with the direction of the marginal effect (ME). Later on, I will describe the relationship between the link function and the distribution of the error terms in random utility models. There the fact that cumulative distribution functions (CDF´s) of continuous random variables, are monotone, continuous and bounded between zero and one, will be useful.

Given the econometric model $E(Y|X) = p(X) = m(X)$, the estimation of the function $p(X)$ differs depending on whether the model is parametric, semiparametric or nonparametric. Parametric estimation assumes that the functional form of $p(X)$ is known and that the conditional probability only depends on the single index $X_i'\beta$.[4] Summarizing, one can say that the parametric approach of modelling binary choice models assumes $p(X_i) = G(X_i'\beta)$, where G is a known function of only one argument $X_i'\beta$. Consequently, the estimation of the model reduces to the estimation of β. The semiparametric approach still assumes that there exists a function $p(X_i) = G(X_i'\beta)$ of a single argument. However, this function is not assumed to be known and is therefore estimated simultaneously with β in a nonparametric fashion. For completeness, the fully nonparametric approach assumes that the function which models $p(X_i)$ can depend in an arbitrary way on all regressors X_j. Therefore, the nonparametric approach generally requires the estimation of a multidimensional function.

The following table gives a summary of the distinction between parametric, semi- and nonparametric approaches to model the conditional probabilities.

Table 1: Characteristics of the parametric, semi- and nonparametric approach

Approach	Model of $p(x)$	known components	unknown components
Parametric	$G(X_i'\beta)$	G	β
Semiparametric	$G(X_i'\beta)$	$G(X_i'\beta)$	G; β
Nonparametric	$m(X_i)$	-	m

Instead of using an ad hoc guess of the form of the link function in the parametric case, Cameron and Trivedi (2005, pp. 475-478) motivate the binary choice model using both the index function model and the additive random utility model. These models suggest a structure to model $p(x)$.

applied work. This is sometimes referred to as the "curse of dimensionality".

[4] Among other things, this means that for any individuals whose combination of X_i´s which result in the same value of $X_i'\beta$ is assumed to have the same probability of $y = 1$. For the LLL estimator the index $X_i'\beta_{X_i}$ varies through X_i and β_{X_i} and Frölich (2004, p. 4) motivates this generalization with the following example considering female labour supply: "The single index restriction imposes that the labour supply effect of, e.g., one versus zero children is identical for all woman for whom the linear combination $X_i'\beta$ has the same value, even if they have very different characteristics".

The index function model assumes that each observation has an unobserved (latent) index y_i^* which can be appropriately explained by the regression

$$y_i^* = X_i'\beta + u \tag{6}$$

Due to the latency of the index we merely observe whether the dependent variable takes the value 0 or 1,

$$y_i = \begin{cases} 1 \text{ if } y_i^* > 0 \\ 0 \text{ if } y_i^* \leq 0 \end{cases} \tag{7}$$

Now noticing that our object of interest is $p(X_i) = p(y = 1|X_i)$, we can rewrite this in terms of y^*. Doing so yields

$$p(X_i) = p(y = 1|X_i) = p(y^* > 0|X_i)$$

$$= p(X_i'\beta + u_i > 0|X_i) = p(u_i > -X_i'\beta|X_i)$$

$$= 1 - p(u_i < -X_i'\beta|X_i) = 1 - F_{u_i|X_i}(-X_i'\beta) \tag{8}$$

This means that the index function model suggests to model $p(X_i) = 1 - F_{u_i|X_i}(-X_i'\beta)$, where $F_{u_i|X_i}$ is the conditional CDF of the error term in the index function model. If one assumes that X_i and u_i are independent and that the individual errors are i.i.d. the underlying distribution function reduces to the unconditional CDF F_u. If the distribution of the error is symmetric, (8) simplifies to $p(y_i = 1|X_i) = F_{u_i|X_i}(X_i'\beta)$. The probit and logit models described in chapter 2.1.2 can be motivated by assuming that u follows a normal distribution (probit) or a logistic one (logit).

The additive random utility model assumes that the choices between the two alternatives are based on utility comparison. Formally, the individual i decides between U_{i0} and U_{i1} depending on their corresponding levels, with U_{i0} and U_{i1} modelled as

$$U_{i0} = X_i'\beta_0 + \epsilon_{i0}$$

$$U_{i1} = X_i'\beta_1 + \epsilon_{i1}$$

Since $y = 1$ is chosen whenever $U_1 > U_0$ it follows that

$$P(y_i = 1|X_i) = P(U_{i1} > U_{i0}|X_i)$$

$$P(X_i'\beta_0 + \epsilon_{i0} > X_i'\beta_1 + \epsilon_{i1}|X_i) = P(\epsilon_{i0} - \epsilon_{i1} > X_i'(\beta_1 - \beta_0)|X_i)$$

$$P(\epsilon_{i0} - \epsilon_{i1} > X_i'(\beta_1 - \beta_0)|X_i) = 1 - F_{\epsilon_{i0}-\epsilon_{i1}|X_i}(X_i'(\beta_1 - \beta_0))$$

Hence, the additive random utility model suggests that $p(X_i) = 1 - F_{\epsilon_{i0}-\epsilon_{i1}|X_i}(X_i'(\beta - \beta_0))$[5], which is again based on the distribution function of the error terms.

As we have seen from the index function model and the additive random utility model there exists a one to one relationship between the distribution of the error terms and the link function. The relation is explicitly given by

$$p(X_i) = G(X_i'\beta) = 1 - F_u(-X_i'\beta) = 1 - F_{\epsilon_0-\epsilon_1}(X_i'(\beta_1 - \beta_0)) \tag{9}$$

If F is symmetric, then the relationship simplifies further:

$$p(X_i) = G(X_i'\beta) = F_u(X_i'\beta) = F_{\epsilon_0-\epsilon_1}(X_i'(\beta_0 - \beta_1)) \tag{10}$$

Summing up the main points of this chapter, the primary goal in the context of binary choice models is captured by the quantity $p(X_i)$, which relates the probability of outcome 1 to observed characteristics. The estimation procedure of $p(X_i)$ can be parametrical, semi- or nonparametrical. The basis of the estimation is the likelihood function $\Psi(p) = \prod_{i=1}^{N} p(X_i)^{y_i}(1 - p(X_i))^{1-y_i}$ whose log-transformation becomes $L(p) = \sum_{i=1}^{N}[y_i ln(p(X_i)) + (1-y_i)ln(1-p(X_i))]$. To obtain some idea which functional form $p(X_i)$ has, one can think of the decisions about y_i descending from the index function model or the random additive utility model. The next part will motivate some quantities of interest, which are used to assess predictive power of the estimators and their appropriateness for causal analysis.

1.2 Quantities of interest

1.2.1 Predictive power

The thesis introduces four measures for the (in-sample) predictive power of the estimators.

i) RMSE: The Root Mean Squared Error (RMSE) is defined as $\sqrt{E((y_i - \hat{E}(y_i|X_i)^2)}$, which equals $\sqrt{E((y_i - \hat{p}(X_i))^2)}$ in binary choice models. Lower values of the RMSE indicate better performance of the estimator.

ii) RMSE80: The RMSE80 is defined as the RMSE by $\sqrt{E((y_i - \hat{E}(y_i|X_i)^2)}$. The difference stems from the fact that after the initial estimation several observations are dropped when calculating the RMSE. The procedure for dropping observations is as follows. First, the observations are demeaned. Then the sum of the squared demeaned dependent variable and the squared demeaned regressors is computed for each observation. Finally, the lower and upper ten percent of the observations (according to this distance measure) are deleted. The

[5] If the errors are i.i.d. and X is independent of ϵ_i then $F_{\epsilon_{i0}-\epsilon_{i1}|X_i}(X_i'(\beta_1 - \beta_0)) = F_{\epsilon_0-\epsilon_1}(X_i'(\beta_1 - \beta_0))$

RMSE80 therefore measures the predictive power in the inner 80% of the sample.

iii) RMSE_p̂: The definition for the RMSE stays the same. However, the predicted value is changed to $\hat{E}(y_i|X_i) = 1$, if $\hat{p} > 0.5$ and $\hat{E}(y_i|X_i) = 0$ otherwise. This measure might be interesting for individuals who have to make a one time decision given the knowledge of $\hat{p}$.[6]

iv) SWP0.95: SWP0.95 is the share of wrong predictions given $\hat{p} > 0.95$**.** This measure assesses the predictive performance at the upper tail. The measure is constructed as follows. First, all individuals for whom $\hat{p}_i < 0.95$ are deleted. Second, all remaining observations are assigned the value one. Third, the difference between one and the realization of y is computed. Then the sum of this differences is divided by the number of individuals whose predicted probability is larger than 0.95. Formally stated:

$$SWP0.95 = \frac{\sum\limits_{i:\hat{p}_i>0.95} (1 - y_i)}{\sum\limits_{i:\hat{p}_i>0.95} (1)} \tag{11}$$

Interest in this statistic can be motivated in the following way. Consider a doctor who wants to calculate the survival probability of a patient for a certain period given some characteristics. He then might use one of the binary choice estimators. After the calculation, the patient receives the pleasant message that his likelihood of surviving a certain period is larger than 95%. The reliability of this information from an individual perspective could be assessed by the measure SWP0.95, which gives in this context the share of dead people, given that their predicted survival probability exceeded 95%.

The following table gives an overview of the measures of predictive power.

Table 2: Overview of the measures for predictive power

Name	Purpose
RMSE	overall in-sample predictive power
RMSE80	predictive power in the inner 80% of the sample
RMSE_p̂	predictive power given zero-one decisions
SWP0.95	predictive power at the upper tail of the index

1.2.2 Causal (marginal) effects

Expected causal effects answer questions of the kind: "What is the expected change in the dependent variable $E(y_i|X)$ if the independent variable x_{ij} is changed, holding all other X_{i_j} constant?" In a binary choice context, the change in the conditional expectation coincides with the change in conditional probabilites $p(X_i)$. Since in the general case, $p(X_i)$ is a multidimensional function $p : \mathbb{R}^K \to [0, 1]$, the marginal effects can be described by the partial

[6] There exist more specialized semiparametric estimators like the maximum score estimator by Manski which have superior properties with respect to in-sample prediction, which are not covered in the thesis.

derivative of $p(X_i)$ with respect to x_{ij} denoted by $\frac{\partial p(X_i)}{\partial x_{ij}} \equiv p_j(X_i)$.[7]Alternatively the marginal effects can be defined via finite differences, i.e. $\frac{p(X_i)-p(X_i-e\cdot h)}{h}$, where e is a vector consisting of zeros and a single one placed such as to give the direction of the derivative. Both methods will be employed in the Monte Carlo study. Here it should be noted that this on its own does not reduce the complexity of the object we want to describe. Still, the partial derivatives themselves are multidimensional functions. To receive results which are easily interpretable the partial derivative is evaluated at several points. As common in the literature the evaluation points chosen are the mean and the first and third quartile of X. Evaluation of the partial derivative at the mean is given by $p_j(n^{-1}\sum X_i)$ and called the marginal effect at the sample average. Generally the evaluation at a quantile is given by $p_j(X_q)$, where X_q denotes the vector of the q-th quantile of the individual regressors. $p_j(X_q)$ is called the marginal effect at the sample quantile q. Finally, one can evaluate the partial derivative at each sample point and then average the effects. The result is called average marginal effect and is given by the formula $n^{-1}\sum_{i=1}^{N} p_j(X_i)$. The following table summarizes the marginal effects, given $G(\cdot)$ as the link function which depends on the single index $X_i'\beta$. Further $\frac{\partial G(X_i'\beta)}{\partial(X_i'\beta)} \equiv g(X_i'\beta)$.

Table 3: Overview of the marginal effects

Sample average ME	ME at sample average	ME at sample quantile
$\frac{1}{N}\sum_{i=1}^{N} g(X_i'\beta)\beta_j$	$g(\frac{1}{N}\sum_{i=1}^{N}(X_i'\beta))\beta_j$	$g(X_q'\beta)\beta_j$

In the case of a symmetric error distribution from an index function model, $\frac{\partial G(X_i'\beta)}{\partial(X_i'\beta)} = \frac{\partial F(X_i'\beta)}{\partial(X_i'\beta)} = f(X_i'\beta)$ where f denotes the probability density function (PDF) of the errors.

It depends on the application which of the marginal effects is of interest. The average marginal effect is a good measure for the overall effect of a policy change which affects all individuals. However as usual with averages, the average marginal effect does not capture that the marginal effects might differ substantially across individuals. The marginal effect at some specific point uses a constructed individual, which generally does not exist in the sample and gives the expected change in y given that X_i is equal to the initial point. The marginal effect at the first (third) quartile gives the effect for low (high) values of X_i. The marginal effect at the median and at the mean describe the effects for representative individuals. It should be noticed that $plim(\frac{1}{N}\sum_{i=1}^{N} g(X_i'\beta)\beta_j)$ and $plim(g(\frac{1}{N}\sum_{i=1}^{N}(X_i'\beta))\beta_j)$ do not coincide. Hence, the marginal effect at the average and the average marginal effect are fundamentally different objects. Their relation depends on the shape of the link function and, given concavity or convexity, Jensens inequality might be exploited (details can be found in the appendix). If the researcher has a real world application at hand, it would be desirable to calculate the marginal effects at each point in the sample and then estimate the distribution of the effects along different dimensions. Furthermore, the researcher could describe the marginal effects for

[7]These marginal effects obviously differ from the coefficients β and due to the dependence of $\frac{\partial p(X_i)}{\partial x_{ij}} = \frac{\partial G(X_i'\beta)}{\partial X_i'\beta}\beta_j$ on G it seems not reasonable to compare the parameter estimates, disregarding the assumed or estimated functional relationship.

different groups. An example to clarify the discussion above would be the marginal effect of smoking on the likelihood of having lung cancer given a certain age. The marginal effect at the mean would take the mean value of smoking (e.g. 20% smoke) and age (mean age of 40) and estimate the effect for a "created" individual who has a value for smoking of 0.2 (which is clearly imaginary) and an age of 40. The average marginal effect would calculate the marginal effects (ideally calculated by finite differences with h=1) for all individuals and then average those effects. As one can see from this example it might be interesting to distinguish young people from old ones. One could then compare the average marginal effect only by considering those individuals which are less than 40 years old with the average marginal effect for those who are older than 40.
The only difference between the marginal effects of the parametric and the semiparametric setup is the fact that G has to be estimated in the semiparametric case. Since monotonicity and differentiability imply that $\frac{\partial G(X_i'\beta)}{\partial X_i'\beta}$ has the same sign for all X_i, one observes the direction of the marginal effects by looking at the coefficients β_j. To assess the quality of the esimators, the thesis compares the four different kinds of marginal effects with the theoretical marginal effects and the estimated marginal effects from the efficient parametic model.

2 The estimators

This chapter describes the most common estimators for the conditional probability in the binary choice literature including some of their properties. It will distinguish between parametric and semiparametric estimators. Since the use of semiparametric estimators is not common for young researchers who apply econometric techniques, I will introduce some ideas of semiparametric estimation, so that the thesis is self contained.

2.1 Parametric estimators

To distinguish the parametric estimators from the semiparametric estimators one can think of the binary choice model as a nonlinear regression model. In the case of parametric estimators the functional relationship between the dependent and the independent is assumed to be known.

2.1.1 Linear probability model

The linear probability model (LPM) is the simplest possible model. It assumes that the link function is linear in X_i. Therefore, the econometric model looks as follows:

$$E(y_i|X_i) = p(X_i) = X_i'\beta \ \forall i \in \{1, ..., N\} \tag{12}$$

Hence, the likelihood function of this model is:

$$L(\beta) = \sum_{i=1}^{N} [y_i ln(X_i'\beta) + (1 - y_i) ln(1 - X_i'\beta)] \tag{13}$$

Maximizing the likelihood yields the closed form solution $\hat{\beta} = (X'X)^{-1}X'Y$.[8] Since $\hat{p}(X_i) = X_i'\hat{\beta}$ there might exist an i where $\hat{p}(X_i) \notin [0,1]$, which implies the possibility of predicted probabilities outside the closed zero-one intervall. This is one reason why the estimator seems problematic. The marginal effects implied by this model are independent of X_i and given by $\frac{\partial p(X_i)}{\partial x_{ij}} = \beta_j$. Interestingly, Cameron and Trivedi (2005, p.471) claim that the OLS estimator "provides a reasonable direct estimate of the sample-average marginal effect on the probability that $y_i = 1$ as x_{ij} changes". This hypothesis was reviewed in the Monte Carlo study and receives support. For the sake completeness, it should be noted that the error terms in the LPM are heteroscedastic by construction, and hence heteroscedasticity adjusted standard errors should be used.

2.1.2 Probit, logit and cauchit models

i) The probit model assumes that the link function is the cumulative distribution function (CDF) of a standard normal distribution. Since the PDF of the normal distribution is symmetric, this model can be justified when the error terms in the index function model follow a standard normal distribution.

$$E(y_i|X_i) = p(X_i) = \Phi(X_i'\beta) \ \forall i \in \{1, ..., N\} \tag{14}$$

The likelihood function of the probit model is

$$L(\beta) = \sum_{i=1}^{N} [y_i ln(\Phi(X_i'\beta)) + (1 - y_i) ln(1 - \Phi(X_i'\beta))], \tag{15}$$

where Φ is the CDF of a standard normal distribution. Since the image of $\Phi(z)$ for $z \in \mathbb{R}$ is $[0, 1]$ the probabilities are bound between zero and one. The marginal effects are given by $\frac{\partial p(X_i)}{\partial x_{ij}} = \phi(X_i'\beta)\beta_j$and therefore depend on X_i.[9] A closed form solution does not exist and hence one has to rely on numerical optimization procedures to solve for the parameter estimates $\hat{\beta}$ which maximize the likelihood.

ii) The logit model has the logistic CDF as its link function. Again, due to symmetry of the logistic PDF it follows that a motivation for the logit model can be logistic distributed errors in the index function model.

$$E(y_i|X_i) = p(X_i) = \Lambda(X_i'\beta) \ \forall i \in \{1, ..., N\} \tag{16}$$

and the likelihood of the logit model is given by

$$L(\beta) = \sum_{i=1}^{N} [y_i ln(\Lambda(X_i'\beta)) + (1 - y_i) ln(1 - \Lambda(X_i'\beta))] \tag{17}$$

[8] $\hat{\beta}$ is equivalent to the ordinary least squares estimator (OLS).
[9] Where $\phi(z)$ denotes the PDF of the standard normal distribution

Λ denotes the CDF of the logistic distribution. The structure of the model is similar to the probit model. $p(x)$ is bound between zero and one and the marginal effects are given by $\frac{\partial p(X_i)}{\partial x_{ij}} = \lambda(X_i'\beta)\beta_j$, where $\lambda(z)$ is the PDF of the logistic distribution.
The results of the marginal effects coming from logit and probit models are in general very similar. An advantage of the logit model is the relatively simple functional form of the distribution function which eases numerical optimization. This becomes important when building more sophisticated econometric models like the local logit model. Frölich (2006, p.6) states that the use of the logistic distribution instead of relying on the cumulative distribution function of a standard normal random variable is due to computational ease. The probit model, on the other hand, might have the appeal that it can be derived by the index function model with normal errors, which have a long tradition in econometrics.

iii) The cauchit model has the CDF of a Cauchy distribution[10] as its link function.

$$E(y_i|X_i) = p(X_i) = C(X_i'\beta) \ \forall i \in \{1, ..., N\} \tag{18}$$

and the likelihood of the cauchit model is given by

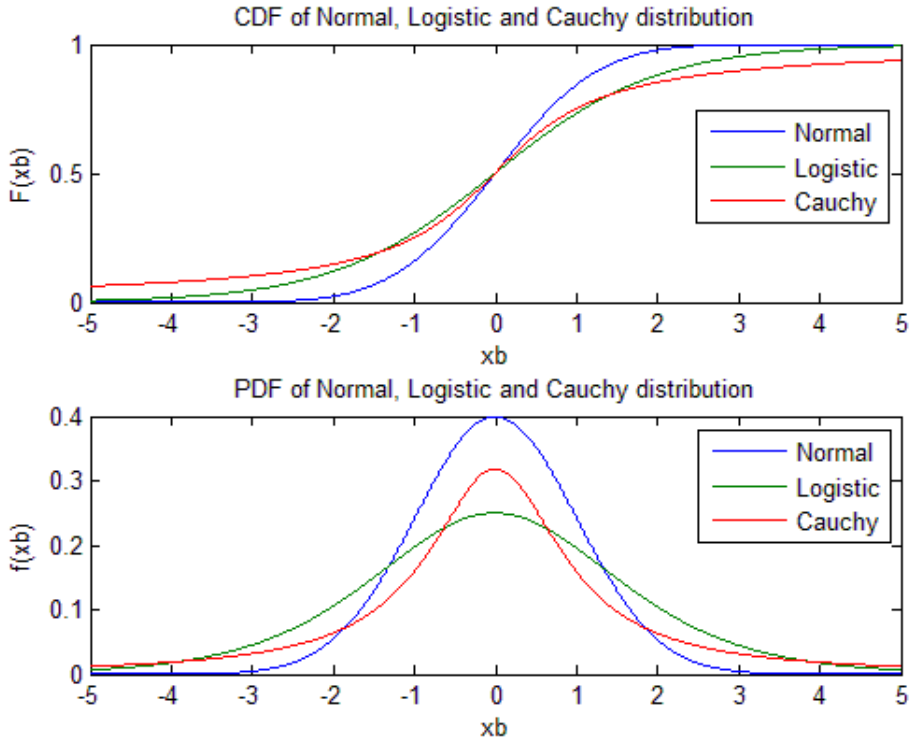
$$L(\beta) = \sum_{i=1}^{N} [y_i ln(C(X_i'\beta)) + (1 - y_i) ln(1 - C(X_i'\beta))] \tag{19}$$

The main purpose of introducing the cauchit model is that it seems suitable to deal with problems such as outliers, through its "fat tails". The marginal effects are given by $\frac{\partial p(X_i)}{\partial x_{ij}} = c(X_i'\beta)\beta_j$.

The following graphs depict the PDF´s and the CDF´s of the standard normal-, logistic-, and Cauchy distribution.

[10]The cauchy distribution is "fat tailed" and even its first moment does not exist.

Figure 1: CDF and PDF of normal-, logistic- and Cauchy distribution



As one can see, the densities are symmetric. Since the marginal effects are given by $\frac{\partial p(X_i)}{\partial x_{ij}} = f(X_i'\beta)\beta_j$, where f denotes the density, one further sees that the marginal effects vary with a maximum at the mean of the single index $X_i'\beta$.

2.1.3 A model using general distributional assumptions

The probit- and the logit model are special cases of the model using general distributional assumptions. The likelihood function of this model is given by

$$L(\beta) = \sum_{i=1}^{N} [y_i ln(1 - F_u(-X_i'\beta)) + (1 - y_i) ln(F_u(-X_i'\beta))] \tag{20}$$

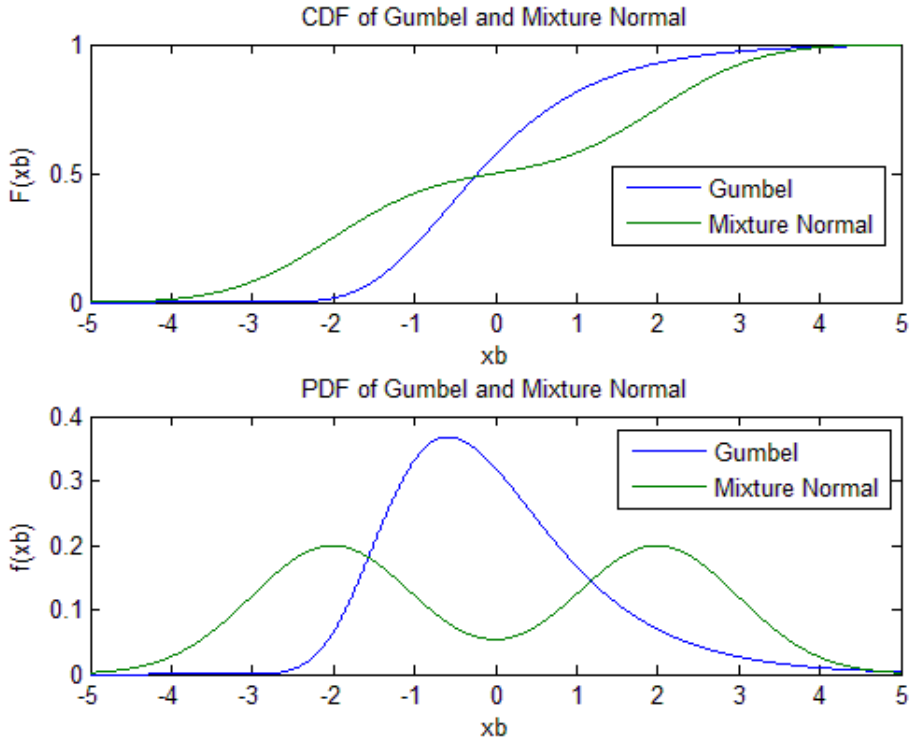
$$L(\beta) = \sum_{i=1}^{N} [y_i ln(G(X_i'\beta)) + (1 - y_i) ln(1 - G(X_i'\beta))] \tag{21}$$

where F_u denotes the error distribution in the index function model. If one knew the true distribution of the error term, using the distribution function as the link function would be generally asymptotically efficient.[11] Clearly, the distribution of the error term is unknown in real world applications. Direct estimation of the distribution of the error terms would require the knowledge of the latent utility index. Since the utilities are not observed, the estimation

[11] This comes from the fact that the estimator is an M-Estimator and under certain regularity conditions is therefore asymptotically efficient, see Amemiya (1985, pp.123-124).

of the error terms distribution seems impossible. However, due to the fact that I create the data generating process (DGP) used in the Monte Carlo study, the model using the correct distributional assumptions will serve as a benchmark. Two further distributions have been used in the Monte Carlo study, namely the Gumbel- and a mixture normal distribution. The PDF´s and CDF´s of the Gumbel- and the considered mixture normal distribution are plotted below. The Gumbel distribution is skewed whereas the mixture normal is multimodal. All random variables are normalized such that they have zero expected value.

Figure 2: CDF and PDF of Gumbel- and mixture normal distribution



2.2 Semiparametric estimators

The terminology for non- and semiparametric estimation varies across authors and instead of defining the terminology rigorously, I will use the terms parametric, semiparametric and nonparametric as follows: "Parametric estimation" refers to the estimation of a finite ordered set of parameters, like in the case of the ordinary least squares, the estimation of the parameter vector β. The term "nonparametric estimation" refers to an infinite-dimensional object like the link function in a binary choice model $\hat{G}(z)$. The term "semiparametric estimation" is used when both a finite dimensional and an infinite dimensional object is estimated, like $\hat{G}(X_i'\hat{\beta})$. Loosely speaking, the semiparametric approach to estimation is a combination of the parametric approach, which is usually concerned with the estimation of parameters, and the nonparametric approach, which is concerned with the estimation of functions.

2.2.1 Introduction to semiparametric estimation

As stated above, semiparametric estimation mixes the parametric- and the nonparametric estimation methods and attempts to combine the benefits of both. The main advantage of the parametric estimation method is its parsimony, whereas the nonparametric approach offers flexibility. To give some intuition behind semiparametric estimation, I will briefly introduce the nonparametric approach and then describe the two semiparametric estimators under consideration. The following introduction of the nonparametric approach is closely related to Cameron and Trivedi (2005, pp. 294-333) and should help readers who are not familiar with nonparametric econometrics in understanding the properties of the estimators presented.

A nonparametric econometric model in a binary choice context can be specified as

$$E(Y|X) \quad = \quad m(X) \tag{22}$$

or in terms of individuals

$$E(y_i|X_i) \quad = \quad m(X_i) \; \forall i \in \{1, .., N\}, \tag{23}$$

where the functional form m is fully unspecified. The whole chapter 2.2.1 will be concerned with the question how a reasonable estimator $\hat{m}$ for m can be formed. To answer this question, it seems helpful to introduce the following concepts and some new terminology. First, we notice that the model implies that we have to estimate m which is a function. A good way to start estimating functions is by estimating the function values for specific points in the domain. A first step could be to just use $\tilde{m}(X_i) = y_i$ as a first guess. Two things are relatively problematic. What happens if we have individuals which have the same regressors but different values for y. An answer to this question could be averaging. Hence $\tilde{m}$ could look as follows: $\tilde{m}(X_i) = \sum_{j=1}^{N} \frac{I(X_j=X_i)}{\sum_{j=1}^{n} I(X_j=X_i)} \cdot y_j$, where I is the indicator function which takes the value one if $X_j = X_i$ and zero otherwise. The second problem would be the estimation of y given that the values for the regressors are not present in the sample. Consequently, the first formula for $\tilde{m}$ would lead to $\hat{E}(y_i|X) = 0 \; \forall X \neq X_i \; \forall i \in \{1, .., n\}$. Hence, the estimator would be extremely nonsmooth. A smoother and relatively general estimator which is able to deliver non-zero values for regressor values not represented in the sample is given by

$$\tilde{m}^{w}(x) = \sum_{i=1}^{N} w_{ni}(x) \cdot y_i \tag{24}$$

This estimator is called local weighted average estimator. The general idea behind this estimator is that we average over y in a specific way. Usually, one assigns lower weights to those values of y where the distance to the initial x is large. The following estimator is a special kind of local weighted average estimator. It is called Nadaraya-Watson (NW) estimator and

is one representative of the class of kernel regression estimators.

$$\hat{m}(x) = \frac{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x}{h}\right) \cdot y_i}{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x}{h}\right)} \tag{25}$$

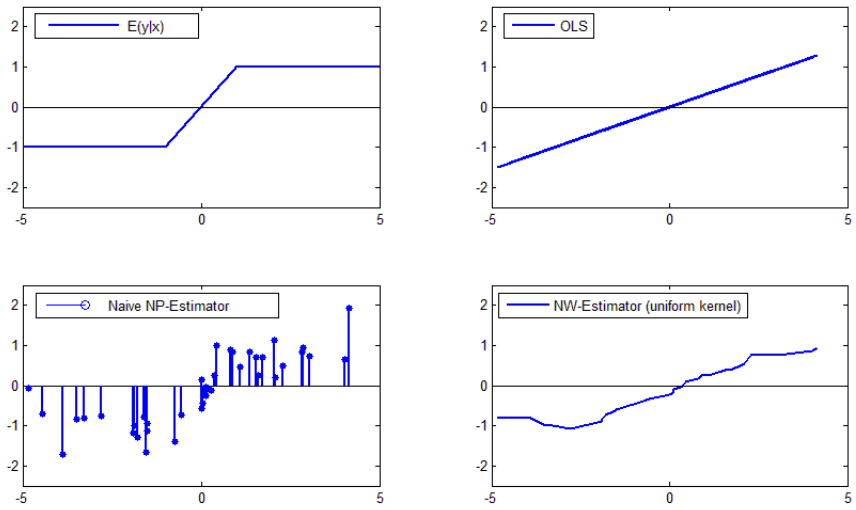
Simple manipulations of (24) reveal the weighting function.

$$\hat{m}(x) = \sum_{i=1}^{N} \frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{N} K\left(\frac{x_i - x}{h}\right)} \cdot y_i = \sum_{i=1}^{N} w_{ni}^{NW}(x) \cdot y_i \tag{26}$$

The weighting function $w_{ni}^{NW}(x) = \frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{N} K\left(\frac{x_i - x}{h}\right)}$ consists of several components. K is a kernel function, h is called the bandwidth and $x_i - x$ is the difference of the sample point to the point of evaluation.

Since these concepts might be new, I give an illustration of the estimators, with OLS as a reference. The DGP is given by $y_i = -I(x_i \leq -1) + x_i I(-1 < x_i < 1) + I(x_i \geq 1) + \epsilon_i$, which is piecewise linear. The regressors x_i are drawn from a normal distribution with mean 0 and variance 4. The error terms are drawn from a normal distribution with mean 0 and variance 0.25. The number of observations is 40 and ϵ and x are independent.

Figure 3: Comparison of estimators: OLS, $\tilde{m}$, $\hat{m}$



As one can see in Figure 3, the assumed linear relation between x and y from the OLS estimator is clearly violated. Further, the relation between the NW estimator and the naive nonparametric estimator becomes apparent. For the NW estimator h is chosen to be Silverman´s plug-in estimate, which will be explained later. Intuitively the NW estimator averages the values of y in the neighbourhood of x. Looking at $x \approx -4$, the values of y result approximately from averaging the y_i´s for x_i lying in the intervall $[-5, -3]$. The length of the

intervall depends on the choice of h.

Since this introductory example is definitely not sufficient to understand the working of kernel regression estimators, the thesis will now discuss the components of those estimators beginning with the kernel function and the bandwidth. The kernel function is discussed via introducing kernel density estimation, which can be seen as a smoothed extension to the well known histogram. In the context of the kernel density estimation, the choice of h plays a crucial role. Therefore the thesis will discuss the choice of h in some detail. After this, the thesis returns to the NW kernel regression estimator and gives a rigorous description of the properties of the estimator.

Kernel density estimation:

The starting point for kernel density estimation is a given sample of data denoted by $\{y_i\}_{i=1}^N$. A first step in approximating the density could be by a nonsmooth estimator like the histogram. A smooth extension is given by a kernel estimate of the density denoted by $\hat{f}(y)$. The estimated density at a point y_0 is given by the formula

$$\hat{f}(y_0) = \frac{1}{Nh} \sum_{i=1}^{N} K(\frac{y_i - y_0}{h}) \tag{27}$$

(27) describes the "height" of the estimated density at an arbitrary point y_0. To construct the value $\hat{f}(y_0)$, the kernel density estimator uses the whole sample and weights the observations according to their difference to the point of interest. The weighting depends on K and h. Following the definition by Cameron and Trivedi (2005, p. 299) "the kernel function $K(\cdot)$ is a continuous function, symmetric around zero, that integrates to unity and satisfies additional boundedness conditions". Examples for kernel functions are given in the following table.

Table 4: Kernel Functions

Name	Kernel function
Uniform	$\frac{1}{2} \cdot I(\lvert\frac{y_i-y_0}{h}\rvert < 1)$
Epanechnikov	$\frac{3}{4} \cdot (1 - [\frac{y_i-y_0}{h}]^2) \cdot I(\lvert\frac{y_i-y_0}{h}\rvert < 1)$
Gaussian	$(2\pi)^{-\frac{1}{2}} exp(\frac{-[\frac{y_i-y_0}{h}]^2}{2})$
Quartic	$\frac{15}{16} \cdot (1 - [\frac{y_i-y_0}{h}]^2)^2 \cdot I(\lvert\frac{y_i-y_0}{h}\rvert < 1)$

Figure 4: Graphs of kernel functions

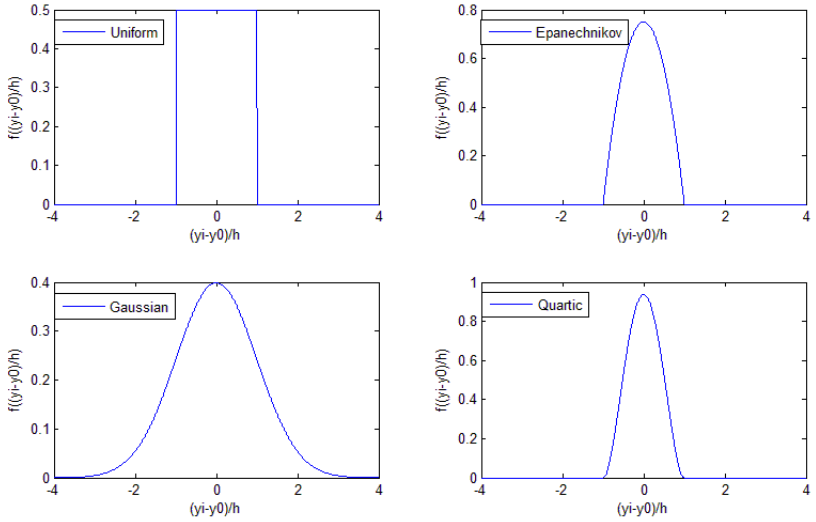
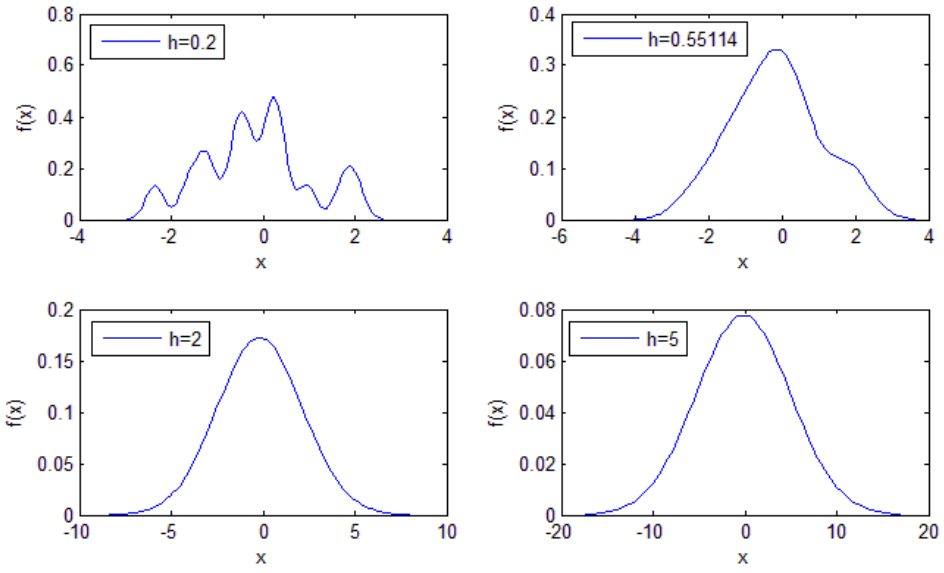


Figure 4 gives an intuition how the sample observations y_i are weighted according to their difference to y_0. If $|y_i - y_0| > h$ the uniform- and Epanechnikov kernel produce a zero weight, whereas the Gaussian kernel gives a positive weight for all observations. The influence of the choice of the kernel function on the estimate of the density is limited. The only thing which is left unclear in formula (27) is the choice of the parameter h. It influences the smoothness and the bias of the density estimator. Maybe the simplest way to illustrate the influence of h on the smoothness is given by little example. The example is constructed by generating 15 observations from a standard normal distribution. The kernel used is Gaussian and the bandwidth h varies across the four plots which are given below.

Figure 5: Kernel density estimate of standard normal data, size 15



The density estimate on the upper right is computed using an optimal bandwidth given the data is normally distributed. One observes that as h increases, the estimator becomes smoother. However the increased smoothness comes at the cost of a higher bias. The bias of an estimator is defined as the difference of the expected value of an estimator to the true value. In the case of the kernel density estimate at y_0, the bias is given by the following formula.

$$b(y_0) = E[\hat{f}(y_0)] - f(y_0) = \frac{1}{2}h^2 f''(y_0) \int z^2 K(z) dz \tag{28}$$

As illustrated in the graphs above, the formula reveals that the bias is increasing in h. Since $\lim_{h \to 0} b(y_0) = 0$, it seems reasonable to choose smaller values for h when larger samples are available. Further, it can be shown that the bias corrected estimator is asymptotically normal. Formally this is given by

$$\sqrt{Nh}(\hat{f}(y_0) - f(y_0) - b(y_0)) \xrightarrow{d} N(0, f(y_0) \int K(z)^2 dz) \tag{29}$$

By dividing the expression above by $\sqrt{Nh}$ one sees that the variance of the estimator $\hat{f}(y_0)$ goes to zero as $Nh \to \infty$.

To sum up, unbiasedness requires $h \to 0$ and consistency $Nh \to \infty$. For the choice of the bandwidth and the kernel one can rely on the mean integrated square error (MISE) as a measure of optimality. The objective function to be minimized is the following:

$$MISE(h) = \int MSE[\hat{f}(y_0)] dy_0 = \int E[(\hat{f}(y_0) - f(y_0))^2] dy_0 \tag{30}$$

where the mean squared error (MSE) is a local measure of performance and is approximately given by $MSE[\hat{f}(y_0)] \backsimeq \frac{1}{Nh} \cdot f(y_0) \int K(z)^2 dz + \{\frac{1}{2}h^2 f''(y_0) \int z^2 K(z) dz\}^2$. The dependence of the MSE on h through the kernel function complicates the minimization. Strictly speaking the optimal kernel and the optimal h can not be chosen independently. Minimization with respect to h yields

$$\frac{dMISE}{dh} = 0 \Longleftrightarrow h^* = \delta(\int f''(y_0)^2 dy_0)^{-0.2} N^{-0.2} \tag{31}$$

where $\delta = \left(\frac{\int K(z)^2 dz}{(\int z^2 K(z) dz)^2}\right)^{0.2}$ and depends on the kernel chosen. Furthermore, it can be shown that the Epanechnikov kernel is optimal. Since h^* depends on the true curvature of the unknown density given by f'', the formula is not directly applicaple.

There are two main approaches how to choose h. The first one relies on plug-in estimates, the second on Cross-Validation (CV).

i) Plug-in estimate: Cameron and Trivedi (2005, p. 304) state that "a plug-in estimate for the bandwidth is a simple formula for h that depends on the sample size N and the sample

standard deviation". Silverman´s plug-in estimate is given by

$$h^* = \delta \left(\frac{3}{8\sqrt{\pi} \cdot min(s^5, iqr^5/1.349)} \right)^{-0.2} N^{-0.2} \tag{32}$$

which is optimal for normally distributed data, since $\int f''(x_0)^2 dx_0 = \frac{3}{8\sqrt{\pi}\sigma^5}$.[12] According to Cameron and Trivedi (2005, p. 304), "these plug-in estimates for h work well in practice, especially for symmetric unimodal densities, even if $f(x)$ is not the normal density. Nonetheless, one should check by using variations such as twice and half the plug-in estimate."

ii) Cross Validation: CV is a data driven approach which chooses h by minimizing a monotone transformation of $\int (\hat{f}(y) - f(y))^2 dy$ which is the integrated squared error (ISE). The mathematics carried out by Pagan and Ullah (1999, p. 51) show that choosing h by minimizing the ISE is equivalent to [13]

$$\hat{h}^*_{CV} = argmin(CV(h)) = argmin(\frac{1}{N^2 h} \sum_i \sum_j K \circ K \left(\frac{x_i - x_j}{h} \right) - \frac{2}{N} \sum_{i=1}^{N} \hat{f}_{-i}(y_i)) \tag{33}$$

Furthermore it can be shown that $\hat{h}^*_{CV} \xrightarrow{p} h_{opt}$, with a rate of convergence of $n^{-\frac{1}{10}}$. This procedure is computationally intensive, and will not be used in the Monte Carlo study. Generally the use of plug-in estimates is weakly inferior to cross validation. Hence, the thesis might underestimate the performance of the semiparametric estimators. Now that the basic ideas of nonparametric density estimation have been introduced we can proceed to a major ingredient of the semiparametric estimators: the kernel regression.

Kernel regression:
As stated in the beginning of chapter 2.2.1 the econometric model is

$$E(y_i|X_i) \quad = \quad m(X_i) \; \forall i \in \{1, .., n\} \tag{34}$$

Leaving out the indices and inserting the definition of $E(y|x) = \int y \cdot f(y|x) dy$ and noting that $f(y|x) = \frac{f(y,x)}{f(x)}$ results in $m(x) = \int y \frac{f(y,x)}{f(x)} dy$ where a natural estimator is given by $\hat{m}(x) = \int y \frac{\hat{f}(y,x)}{\hat{f}(x)} dy$. Pagan and Ullah (1999, pp.83-84) show that, given a symmetric kernel, $\hat{m}$ can be expressed as follows.

$$\hat{m}(x_0) = \frac{\frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right) \cdot y_i}{\frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)} \tag{35}$$

Therefore, the kernel regression estimator is simply obtained by combining the kernel density

[12] s denotes the sample standard devation and iqr the interquartile range, which is robust to outliers.

[13] $K \circ K$ is the convolution of the kernel functions and $\hat{f}_{-i}(y_i)$ is the leave one out kernel density estimator given by $\hat{f}_{-i}(y_i) = \frac{1}{Nh} \sum_{j \neq i}^{N} K(\frac{y_j - y_i}{h})$.

estimator with the data on the dependent variable. $\hat{m}(x)$ is biased where the bias is given by

$$b(x_0) = h^2(m^{'}(x_0)\frac{f^{'}(x_0)}{f(x_0)} + \frac{1}{2}m^{''}(x_0)) \int z^2 K(z)dz \tag{36}$$

and the bias corrected kernel regressor is asymptotically normal:[14]

$$\sqrt{Nh}(\hat{m}(x_0) - m(x_0) - b(x_0) \xrightarrow{d} N(0, \frac{\sigma_\epsilon^2}{f(x_0)} \int K(z)^2 dz) \tag{37}$$

Plug-in estimates for h as in the case of kernel density estimation can be used. In the case of multiple regressors I used plug-in estimates for the univariate regressors seperately which might be justified by the fact that most Monte Carlo setups used regressors drawn from independent normals. Hence the gain of cross validation could be limited[15]. As in the kernel density estimation, cross validation procedures could be used to determine the optimal h^*_{CV}. This procedure is computationaly intensive but the choice of h is more robust to deviations from normality. The criterion to be minimized is

$$CV(h) = \sum_{i=1}^{N} (y_i - \hat{m}_{-i}(x_i))^2 \pi(x_i) \tag{38}$$

As Cameron and Trivedi (2005, p. 315) state "the weights $\pi(x_i)$ are introduced to downweight the end points" and $\hat{m}_{-i}$ is the leave one out estimator given by:

$$\hat{m}_{-i}(x_i) = \frac{\sum_{j \neq i} K\left(\frac{x_j - x_i}{h}\right) y_j}{\sum_{j \neq i} K\left(\frac{x_j - x_i}{h}\right)} \tag{39}$$

Further "it can be shown that $y_i - \hat{m}_{-i}(x_i) = \frac{y - \hat{m}(x_i)}{1 - K\left(\frac{x_i - x_i}{h}\right) / \sum_j K\left(\frac{x_j - x_i}{h}\right)}$ so that for each value of h cross validation requires only one computation of $\hat{m}(x_i)$, $i = 1, ..., N$", Cameron Trivedi (2005, p.315). However, as one will see in the section presenting the local likelihood logit estimator, this would at least require an additional minimization at each sample point, given an initial estimate of the parametric component, which would assume that the parametric component was chosen correctly.

Two further refinements of the Nadaraya-Watson estimator could be considered. The first one is concerned with outliers, the second one with the problem of values near zero in the denominator of $\hat{m}(x_0) = \frac{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right) \cdot y_i}{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)}$. The discussion is again not innovative and resembles the discussions in standard textbooks like Pagan and Ullah (1999).

To deal with the sensitivity of the kernel regression to outliers, one can use leave one out estimators. To illustrate this, assume a dataset contains an extreme outlier (x_o, y_o).

[14] Where σ_ϵ^2 is the variance of the error term resulting from the model equivalent to (34), which is given by $y_i = m(X_i) + \epsilon_i$, $E(\epsilon_i | X_i) = 0 \ \forall i \in \{1, .., n\}$

[15] It might even be that the estimation of the optimal bandwidth might lead to additional noise as discussed in Frölich (2006, p. 7).

For points sufficiently far apart from the rest $\sum_{j\neq o} K\left(\frac{x_j - x_o}{h}\right) \approx \sum_{j\neq o} 0 = 0$. Hence $\hat{m}(x_o) \approx \frac{K\left(\frac{x_o - x_o}{h}\right)\cdot y_o}{0+K\left(\frac{x_o - x_o}{h}\right)} = y_o$. Thus, one value determines the whole estimate. To overcome this problem, one can use the leave one out estimator. $\hat{m}_{-o}(x_o) = \frac{\sum_{j\neq o} K\left(\frac{x_j - x_o}{h}\right) y_j}{\sum_{j\neq o} K\left(\frac{x_j - x_o}{h}\right)}$. This unfortunately results in $\hat{m}_{-o}(x_o) \approx \frac{0}{0}$, which directly leads us to the procedure of trimming. Trimming is concerned with the problem of denominators near zero. As Cameron and Trivedi (2005, p. 317) state, trimming greatly downweights all points with $\hat{f}(x_i) < b$ which are the points where the kernel density estimator predicts smaller probability mass than b, where b should be small and decreasing with N.[16]

To bridge the gap to binary choice models, equation (40) introduces the semiparametric single index model.

$$E(y_i|X_i) = m(X_i'\beta) \tag{40}$$

where m is an unknown function to be estimated via kernel regression and β are the unknown parameters which will be jointly maximized via M-estimation. One point is worth adding. Horrowitz (2009, p. 13-14) states that in semiparametric models, β is identified only if a location and scale normalization is introduced. This can be seen by the following argument. Imagine that the true model is given by $E(y_i|X_i) = G(X_i'\beta)$, then one can find a $G^*(\alpha + \delta X_i'\beta) = G(X_i'\beta)$ via an appropriate concatenation of two functions. Since the code used for the Klein and Spady estimator provided bei Yingying Dong[17] normalizes the last coefficient to one the relation between the initial function $m(X_i'\beta)$ and the normalized function $m(v(X_i,\theta))$ is as follows: $X_i'\beta = 1\cdot\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k = a + b\cdot v(X_i,\theta)$, where $a = \beta_1$, $b = \beta_k$ and $v(X_i,\theta) = x_{2i}\theta_1 + .. + x_{k-1,i}\theta_{k-2} + x_k$, with $\theta_1 = \beta_2/\beta_k$,..., $\theta_{k-2} = \beta_{k-1}/\beta_k$.[18]

2.2.2 Klein and Spady

In what follows, I will discuss the main properties of the estimator by Klein and Spady, like the likelihood function and its asymptotic efficiency. There exist several implementations which differ in the degree of complexity. Three of them will be discussed. After the technical introduction to the estimator, I will present the details of the implementation procedure.

The original likelihood function suggested by Klein and Spady is given below. The discussion mainly summarizes what is stated in Pagan and Ullah (1999, pp.284).[19]

$$L = \sum_{i=1}^{N} \zeta_i(1-y_i) ln(1-\hat{m}_{-i}(v(x_i,\theta))) + \sum_{i=1}^{N} \zeta_i y_i ln(\hat{m}_{-i}(v(x_i,\theta))) \tag{41}$$

[16] As stated several times before, the refinement of the semiparametric estimators was not a key part of the thesis. However some robustness checks were conducted, as one will see in the following chapters. Admittedly the performance of the semiparametric estimators might improve through refinements.

[17] The code for Klein and Spady´s estimator is downloaded from http://www.yingyingdong.com/Codes/KleinSpady.m.txt, last accesed on 09.01.2012

[18] In the Monte Carlo setup the regressors do not contain a constant. Hence a scale normalization is not needed.

[19] With minor modifications with respect to the estimator of m. In my understanding the original article by Klein and Spady (1993, p. 394), suggests the use of a leave one out estimator.

$$\hat{m}_{-i} = \frac{\sum_{j \neq i}^{N} y_j K((v_j - v_i)/h)}{\sum_{j \neq i}^{N} K((v_j - v_i)/h)} \tag{42}$$

$$E(y_i|X_i) = m(v(x_i, \theta)) \; \forall i \in \{1, ..., N\} \tag{43}$$

As in the parametric case, the likelihood function is given by the product of Bernoulli distributions. The only difference to the parametric likelihoods is that the functional relation between the parameter p of the Bernoulli distribution and the regressors is left, up to the index, unspecified.[20] Hence the econometric model for the conditional expectation is given by (43). To obtain the functional relationship between the single index $v(x_i, \theta)$ and p, a kernel regression of y on the linear index $v(x_i, \theta)$ is performed. The estimator $\hat{\theta}$ is then found by maximizing (41).

Two parts of (41) seem non-standard. First, ζ_i denotes a trimming function, which down-weights or even discards those observations where the kernel density estimate of the single index is below a given level, $\hat{f}(v(x_i, \theta)) < b$. Second, $\hat{m}_{-i}$ is a leave one out kernel estimator. Pagan and Ullah (1999, pp. 284-285) state, that the trimming function is necessary for the derivation of the theoretical results, like asymptotic normality of $\hat{\theta}$. However, Klein and Spady (1993, p. 406) claim that "there is a wide range of trimming specifications that have almost no effect on the estimates. Moreover, the estimate obtained without any trimming performed quite similar to that under the trimming that we employed." To motivate why Klein and Spady´s estimator behaves quite differently than the usual parametric ones, despite the similarity in the likelihood function, it is useful to compare the first order conditions (FOC) of the estimators.

To make the main points clear, we compare a symmetric parametric model, with a version of Klein and Spady´s estimator which is untrimmed and uses an ordinary kernel regression.

Table 5: First order conditions: general parametric model vs. Klein and Spady´s

Likelihood Parametric	$L(\beta) = \sum_{i=1}^{N} [y_i ln(G(X_i'\beta)) + (1 - y_i) ln(1 - G(X_i'\beta))]$
FOC (for β_j) Parametric	$\sum_{i=1}^{N} [\frac{G'}{G} y_i x_{ij} - \frac{G'}{1-G}(1 - y_i) x_{ij}] = 0$
Likelihood Klein Spady	$L(\theta) = \sum_{i=1}^{N} [y_i ln(\frac{\sum_{j=1}^{N} y_j K((v_j - v_i)/h)}{\sum_{j=1}^{N} K((v_j - v_i)/h)}) + (1 - y_i) ln(1 - \frac{\sum_{j=1}^{N} y_j K((v_j - v_i)/h)}{\sum_{j=1}^{N} K((v_j - v_i)/h)})]$
FOC (for θ_j) Klein Spady	$\sum_{i=1}^{N} \left[\frac{y_i}{\hat{m}_i} \frac{d\left(\frac{\sum_{j=1}^{N} y_j K((v_j - v_i)/h)}{\sum_{j=1}^{N} K((v_j - v_i)/h)}\right)}{d\theta_j} - \frac{(1-y_i)}{1-\hat{m}_i} \frac{d\left(\frac{\sum_{j=1}^{N} y_j K((v_j - v_i)/h)}{\sum_{j=1}^{N} K((v_j - v_i)/h)}\right)}{d\theta_j} \right] = 0$

Considering the FOCs might help understanding the complicated dependence structure of the likelihood function on θ. The main difference is that the change in β_j in the parametric

[20] In other words: it is only specified that the functional relation depends on a single index, which will be assumed to be linear.

models only alter the evaluation points of $\frac{G'}{G}$ or $\frac{G'}{1-G}$, whereas changes in θ_j lead to changes in the whole estimate of $\hat{m}$ which are additionally weighted by the change in $\frac{\sum y_j K((v_j - v_i)/h)}{\sum_{j=1}^{N} K((v_j - v_i)/h)}$.[21] The reason for stressing the difference between the estimators will become clear when considering the difference between the logit and local likelihood logit estimator. By focusing on the first order conditions, I came to the conclusion that the local likelihood logit estimator is a local logit estimator in the sense, that it only uses a subsample of the data near the point of evaluation.[22] Such an interpretation does not exist for the estimator of Klein and Spady.

Next we turn to the properties of the estimator by Klein and Spady. The estimator is consistent and $\sqrt{n}$ asymptotically normal.[23] Furthermore, Klein and Spady (1993, p. 405, Theorem 5) state and prove the asymptotic efficiency of the estimator. Klein and Spady´s estimator is optimal in the sense that it has the lowest possible asymptotic variance in the class of consistent semiparametric binary choice estimators. This means that the estimator attains the semiparametric efficiency bound specified in Chamberlain (1986) and Cosslet (1987), which is a similar concept as the more popular Cramer Rao bound for parametric models. The Cramer Rao parametric efficiency bound is attained by those parametric maximum likelihood estimators which fulfill some regularity conditions and have a properly specified likelihood. Due to the fact that the Cramer Rao bound is below the semiparametric asymptotic efficiency bound the following corollary should hold.

Corollary 1: Given a linear single index and a correctly specified link function, the following performance ranking should be resembled in the Monte Carlo study in large samples. The parametric estimator with true link function is more efficient than Klein and Spady´s, which in turn is more efficient than the local likelihood logit estimator.

Further, since Klein and Spady´s estimator is consistent, theory suggests that it performs better than misspecified parametric models.

Implementation methods of Klein and Spady´s estimator:

As stated before, the implementation of Klein and Spady´s estimator varies across authors. Three implementations were considered. Two implementations were programmed by the author and one is taken from Yingying Dong.[24] Due to the fact that the code written by the author took about 40 minutes to find the optimal $\hat{\theta}$´s, the code by Yingying Dong is used in the Monte Carlo study. The different kinds of implementations and the estimation of marginal effects are discussed in the following. The main procedure is as follows. First, we obtain $\hat{\theta}$ (three different methods for that), than estimate $\hat{m}$´s via kernel regression and finally we estimate the marginal effects of interest via finite differences.

[21] Where the change comes through the depndence of v_j and v_i on θ_j

[22] As discussed later, this interpretation of only using a subsample is valid if a uniform kernel is used. However a similar interpretaion should be valid for other kernels

[23] Details and proofs are provided in Klein and Spady (1993).

[24] The code for Klein and Spady´s estimator is downloaded from http://www.yingyingdong.com/Codes/KleinSpady.m.txt, last accesed on 09.01.2012

Implementation I:

$\hat{\theta} = argmax \left(\sum_{i=1}^{N} \zeta_i (1 - y_i) ln(1 - \hat{m}_{-i}(v(x_i, \theta))) + \sum_{i=1}^{N} \zeta_i y_i ln(\hat{m}_{-i}(v(x_i, \theta))) \right)$, where $\zeta_i = I(\hat{f}_i(v(x_i, \tilde{\theta})) > b)$. Hence, the trimming function discards the observations where the estimated density of the index $\hat{f}_i(v(x_i, \tilde{\theta}))$ is smaller than b. For the leave-one-out kernel regression and kernel density estimation, the Epanechnikov kernel and Silverman´s plug-in estimate for h are used.

Implementation II:

$\hat{\theta} = argmax \left(\sum_{i \in N_n^*} (1 - y_i) ln(1 - \hat{m}_{-i}(v(x_i, \theta))) + \sum_{i \in N_n^*} y_i ln(\hat{m}_{-i}(v(x_i, \theta))) \right)$, where the trimming function is 1, and N_n^* denotes the set N where the n most extreme observations are discarded. The leave one out kernel regression uses the Epanechnikov kernel and Silverman´s plug-in estimate of h.

Implementation III:

$\hat{\theta} = argmax \left(\sum_{i=1}^{N} (1 - y_i) ln(1 - \hat{m}_i(v(x_i, \theta))) + \sum_{i=1}^{N} y_i ln(\hat{m}_i(v(x_i, \theta))) \right)$, which is the default implementation in Yingying Dong´s code. The ordinary kernel regression uses a quartic kernel and Silverman´s plug-in estimate for h. Furthermore this implementation does not use any trimming.

The results of a naive comparison with 20 draws with 100 observations and three regressors are the following. The values displayed in Table 6 are the mean of the percentage deviations of the first two methods with the third method. The standard deviations are given in parentheses.[25]

Table 6: Comparison: implementation methods for KS´s estimator

	β_1	β_2	Comment
$\frac{\beta_{qIII} - \beta_{qI}}{\beta_{qI}}$	-0.02 (0.07)	-0.02 (0.07)	Trimming $b = 0.002$
$\frac{\beta_{qIII} - \beta_{qI}}{\beta_{qI}}$	-0.01 (0.09)	-0.03 (0.09)	Trimming $b = 0.05$
$\frac{\beta_{qIII} - \beta_{qII}}{\beta_{qII}}$	-0.03 (0.09)	-0.02 (0.09)	2% of sample cut

On average, the deviations lie between 1% and 3%, which can be regarded as minor. However, looking at the standard deviation reveals that the individual deviations might be substantial. Hence, further research could undertake a more extensive comparison between the different implementation methods.

The Monte Carlo study will use, mainly because of computational effort, implementation method III. To obtain estimates of y, I first constructed the estimate of the single index given

[25]The regressors and the error terms are drawn from independent standard normal distributions. The trimming parameter $\tilde{\theta}$ equals the probit estimate $\hat{\beta}_{Pr}$. An extensive comparison was not possible due to time constraints, partially due to the fact that the estimators programmed by myself where computationally so burdensome, that a single optimization required 40 min (on a dual core processor with 4 GB RAM).

by $X'\hat{\theta}$. The estimate for the single index is then used as an input in a simple kernel regression using a quartic kernel and Silverman´s plug-in estimate. The outcome of this estimation is $\hat{y}$. The marginal effects are then computed by finite differences. More exactly for the estimation of the partial marginal effect of variable k at a given point X_0 I estimated $\hat{y}_0$. Then I fixed the value of the regressor at variable k and searched the value of the nearest not identical value of variable k in the sample. Given the nearest value of variable k I held all other variables constant and estimated $\hat{y}_{nn}$. Then I divided by the difference between the initial value of the regressor at variable k and the value of the k-th regressor at the "nearest" sample point. Formally:

$$\hat{y}_0 = \frac{\sum_j^N y_j K((X_j\hat{\theta} - X_0\hat{\theta})/h)}{\sum_j^N K((X_j\hat{\theta} - X_0\hat{\theta})/h)} \tag{44}$$

$$ME_{FD}(X_0) = \frac{y_0 - y_{nn}}{x_{0k} - x_{nnk}} \tag{45}$$

The thesis will now describe the second semiparametric estimator under consideration, the local likelihood logit estimator by Frölich (2006).

2.2.3 Local likelihood logit

The main part of the description below relies on the primary source by Frölich (2006). The local likelihood logit estimator is defined by its local likelihood function.

$$L(x_0) = \sum_{i=1}^{N}(y_i ln\left(\frac{1}{1+e^{-X_i'\theta_{x_0}}}\right) + (1-y_i) ln\left(\frac{1}{1+e^{X_i'\theta_{x_0}}}\right)) \cdot K_H\left(X_i - X_0\right) \tag{46}$$

$$K_{h,\delta,\lambda}(X_i - X_0) = \prod_{q=1}^{q1}\kappa(\frac{X_{q,i} - X_{q,0}}{h}) \cdot \prod_{q=q_1+1}^{q_2}\delta^{|X_{q,i}-X_{q,0}|} \cdot \prod_{q=q_2+1}^{Q}\lambda^{1(X_{q,i}\neq X_{q,0})} \tag{47}$$

$$E(y|X_0) = \frac{1}{1+e^{-X_0'\theta_{x_0}}} \; \forall i \in \{1, ..., N\} \tag{48}$$

Equation (47) describes the kernel function. Frölich (2006) uses a product kernel, which differs across three types of regressors. The types of regressors are: continuous regressors and discrete regressors with and without natural ordering. The first $q1$ regressors are the continuous ones, the regressors from $q1 + 1$ to $q2$ are the discrete regressors with natural ordering, and the remaining ones are those without ordering. Unlike in the original paper where the bandwidth h, δ and λ are chosen by cross validation, the thesis uses different plug-in estimates.[26] There are several differences to Klein and Spady´s estimator. First the functional form of the link

[26]Frölich (2006) uses the following cross validation criterion which is minimized: $CV_{LS} = \sum[Y_i - g(X_i, \hat{\theta}_{-X_i|h,\delta,\lambda})]$, where $\hat{\theta}_{-X_i|h,\delta,\lambda}$ is the leave one out coefficient estimate. In my understanding this requires N additional optimizations for each individual, which would lead to $(N^2 - N)$ additional optimizations in each loop of the Monte Carlo study. Clearly, as stated above the use of cross validation theoretically enhances the performance of the LLL estimator, and hence there is room for improvements of the local likelihood logit estimator.

function is modelled directly and is given by a logistic CDF. Second, the conditional mean function is defined only locally. Hence, θ is allowed to vary across observations. Third, the likelihood function is defined locally and the estimation procedure for θ is done locally at x_0 via weighting observations using a kernel function. For comparison, Klein and Spady´s likelihood function and the parameter vector θ are defined globally. The link function is as well globally defined, but left unspecified. Table 7 captures the main characteristics of the two semiparametric estimators under consideration.

Table 7: Comparison: Klein and Spady´s- and local likelihood logit estimator

Characteristica	KS	LLL
Likelihood:	Defined globally	Defined locally
Link function:	Defined globally, unspecified	Defined Locally, given by Logistic CDF
Conditional mean:	Defined globally, unspecified	Defined Locally, given by Logistic CDF

Further it is worth looking at the first order conditions, which are displayed in Table 8.

Table 8: First order conditions: general parametric model vs. local likelihood logit

Likelihood Parametric	$L(\beta) = \sum_{i=1}^{N} [y_i ln(G(X_i'\beta)) + (1-y_i) ln(1-G(X_i'\beta))]$
FOC (for β_j) Parametric	$\sum_{i=1}^{N} [\frac{G'}{G} y_i x_{ij} - \frac{G'}{1-G}(1-y_i)x_{ij}] = 0$
Likelihood LLL	$L(\theta) = \sum_{i=1}^{N} [y_i ln(G(X_i'\beta)) + (1-y_i) ln(1-G(X_i'\beta))] \cdot K(X_i - X_0)$
FOC (for θ_j) LLL	$\sum_{i=1}^{N} \left[\frac{G'}{G} y_i x_{ij} - \frac{G'}{1-G}(1-y_i)x_{ij}\right] \cdot K(X_i - X_0) = 0$

As can be already seen from the likelihood function, the similarities to the logit estimator are huge. The easiest way to imagine the effect of the kernel function is by considering a product kernel formed by uniform kernels. Given a uniform kernel for the estimation of $\hat{\theta}_{X_0}$, all observations more distant than h from X_0 are discarded. This means that the likelihood function and the j-th first order condition look as follows.

$$L(x_0) = \sum_{i \in N^*} (y_i ln\left(\frac{1}{1+e^{-X_i'\theta_{x_0}}}\right) + (1-y_i) ln\left(\frac{1}{1+e^{X_i'\theta_{x_0}}}\right)),\ N^* = \{i : (x_{ik} - x_{0k}) < h_k\ \forall k\} \tag{49}$$

$$\sum_{i \in N^*} \left[\frac{G'}{G} y_i x_{ij} - \frac{G'}{1-G}(1-y_i)x_{ij}\right] = 0,\ N^* = \{i : (x_{ik} - x_{0k}) < h_k\ \forall k\} \tag{50}$$

which just means that those parts of the sample are discarded where the distance of the individual regressors is higher than the bandwidth.

Due to the fact that $E(y|X_0) = \frac{1}{1+e^{-X_0'\theta_{x_0}}}$, one can estimate $\hat{y} = \frac{1}{1+e^{-X_0'\hat{\theta}_{x_0}}}$. For the

estimation of the marginal effects, two methods were considered. First, I used the partial derivative of $E(y|X)$ and second finite differences.

i) $\frac{\partial E(y|X_0)}{\partial x_j} = pdf_{logistic}(X_0\theta_{x_0}) \cdot \theta_{jx_0}$ This method appears problematic in the following sense. The model for the conditional expectation of $E(y|X)$ assumes that variation in X affects $E(y|X)$ through two channels. First $E(y|X)$ changes due to the change in X and second through changes in θ_x. Method one does not account for any changes in θ_x and therefore could be seen as implausible.

ii) $ME_{FD}(X_0) = \frac{(y|X_0,\theta_{X_0})-(y|X_T,\theta_{X_T})}{x_{0k}-x_{Tk}}$**.** The second method is similar to the one used for Klein and Spady´s estimator. The only difference is that the distance between x_{0k} and x_{Tk} is fixed at 0.02. X_T is a transformation of X where all elements except the k-th are the same and the k-th element itself is adjusted by 0.02, formally: $x_{Tk} = x_{0k} - 0.02$. The fixed choice of the distance has the disadvantage that the marginal effects should not be interpreted as derivatives even in large samples. The advantage consists in the fact that the denominator does not come too close to zero and the predicted values generally have a difference which is measured larger than zero by the computer. Furthermore, the "marginal" effects have the clear interpretation of varying x_{ik} by 0.02 units.[27]

3 The Monte Carlo study

3.1 Monte Carlo studies in general

Since the objective of the thesis is to compare different properties of estimators, one needs a method which allows such comparisons. Instead of comparing the estimators on theoretical basis, I will conduct several Monte Carlo studies. The structure of the Monte Carlo study is the following. First, I will generate data from a known true data generating process.[28] After the data is generated, several estimators are calculated. After obtaining the estimators of interest, several quantities (RMSE´s and ME´s) are calculated and compared with those of the true DGP. This procedure is iterated several times. In the following, I will describe the setup of the Monte Carlo study.

The Monte Carlo study is done in three steps and then iterated R times.

i) Data is generated from a DGP:

A researcher who conducts an econometric study in a binary choice context usually has the following data at hand: $\{y_i, X_i\}_{i=1}^N$ where y_i is the dependent and X_i are the independent variables. In a first step, the Monte Carlo study assumes a specific DGP and generates the data. For interpretational convenience, the data is generated using an index function model. This has been implemented as follows.

a) Draw N-times X_i´s and u_i´s from some random number generator.

b) Choose a function which links the X_i´s and u_i´s to the latent index y_i^*. Usually this is done in a linear fashion, i.e. $y_i^* = X_i\beta + u_i$. Hence, in the linear case one has to fix the true

[27] 0.02 units usually correspond to a change which is 2% of the standard deviation of the x_{ik}´s.

[28] These data will come from random number generators in MATLAB.

coefficient vector β.

c) Given y_i^*, one generates the y_i using the following rule: for all i with $y_i^* > 0$, choose $y_i = 1$. For the remaining i ´s choose $y_i = 0$. Notice, that at this point we know the quantities $\{y_i, y_i^*, X_i, \beta, u_i\}_{i=1}^N$. Since the researcher only knows $\{y_i, X_i\}_{i=1}^N$ these are the only information which are used in the next step.

ii) Estimation:

As stated in chapter 1 we are interested in different properties of the estimators. Usually statistical properties of interest are the following. Is an estimator unbiased? Is it consistent? And given that it is both, which estimator is most efficient. Still the question remains: "what is the object of interest?"

In the linear regression framework the standard answer would be β. However, in the binary choice context an interpretation of β_j would be that it corresponds to the expected marginal effect of an increase in x_j on the latent index (which can represent utility) and is hard to interpret. Hence, the interest of the researcher in the parameter value β is assumed to be limited.[29] It seems plausible that the following two quantities are of major interest in the binary choice context: $\hat{y}_i$ the predicted value for the likelihood that an individual has $y = 1$ and the $\hat{ME}$ ´s, which are the estimators of the marginal effects discussed in chapter 1. For the estimator of $\hat{y}$ the RMSE is used as the main of performance measure. For the $\hat{ME}$´s the property of unbiasedness is considered via looking at small samples. Consistency is informally checked by looking at large samples and the behaviour of the variance of the $\hat{ME}$´s. Efficiency is considered via the comparison of the variance of those estimators which are consistent.

a) For each of the seven estimators the quantities of interest are calculated. The quantities of interest are RMSE, RMSE80, RMSEs0_5, SWP0_95, for the performance of the estimators $\hat{y}$, and the marginal effects at the first- and third quartile as well as the marginal effects at the average and the average marginal effects (for details see chapter 1).

b) Given the true DGP, the true marginal effects are calculated in each iteration by replacing the realizations of X_i in the corresponding true functional form of the marginal effects.

iii) Repetition:

After the computation of the quantities of interest a new set of random variables, i.e. $\{X_i, u_i\}_{i=1}^N$, is drawn and step i) and ii) are repeated R times.

iv) Summary of results:

After the R repetitions the results of the Monte Carlo study are summarized and presented.

[29] Furthermore it is relatively obvious that the parameter estimate will depart substantially from the true parameter, when the functional form is misspecified. Still it might be the case that the marginal effect on the probability of $y = 1$ (the observed variable) might be close to the truth, even under misspecification of the functional form.

3.2 The different Monte Carlo setups

Since the results of the Monte Carlo study are specific to the assumed DGP, the DGP is varied across many dimensions. Ten different setups are considered. There is a clear division between the 10 setups. The setups S1 i) - S1 vii) share the following property. Asymptotic theory predicts, that under each of the setups one parametric estimator is efficient and Klein and Spady´s estimator is consistent. The setups S2 i) - S2 iii) share the following. Theory predicts, that none of the estimators under consideration is consistent. The main point of this analysis is how severe the deviations from the truth are (for example, if the signs of the marginal effects change). The following table gives a summary of the different setups.

Table 9: Summary of the Monte Carlo setups

Setup	DGP	Main concern
S1: i)	$y^* = \sum_{j=1}^{3} \beta_j x_j + u$; u is normal	Performance under standard assumptions
S1: ii)	$y^* = \sum_{j=1}^{3} \beta_j x_j + u$; u is logistic	Performance under standard assumptions
S1: iii)	$y^* = \sum_{j=1}^{3} \beta_j x_j + u$; u is Cauchy	Performance under fat tails
S1: iv)	$y^* = \sum_{j=1}^{3} \beta_j x_j + u$; u is Gumbel	Performance under skewness
S1: v)	$y^* = \sum_{j=1}^{3} \beta_j x_j + u$; u is mixture normal	Performance under multimodality
S1: vi)	$y^* = \sum_{j=1}^{6} \beta_j x_j + u$; u is normal	Performance with many regressors
S1: vii)	$y^* = \sum_{j=1}^{3} \beta_j x_j + u$; u has outliers	Performance under presence of "outliers"
S2: i)	$y^* = \sum_{j=1}^{2} \beta_j x_j + ln(x_3) + u$; u is normal	Behaviour under misspecified index
S2: ii)	$y^* = \left(\sum_{j=1}^{3} \beta_j x_j\right)^3 + u$; u is normal	Behaviour under misspecified index
S2: iii)	$y^* = \sum_{j=1}^{4} \beta_j x_j + u$; u is normal	Behaviour under omitted variables

In the following, I will give a more detailed description of the different Monte Carlo setups. If not mentioned differently, the regressors X_i are drawn from the standard normal distribution. Further, the marginal effects are calculated for the last observed variable and the number of repetitions was 200. The number of observations was element of the set Nobs $= \{50, 100, 250, 500, 750, 1000\}$.

Setup 1 i) is the setup which is often implicitly assumed, when a researcher uses a probit model. The DGP is given by $y^* = \sum_{j=1}^{3} \beta_j x_j + u$, where the true β coefficient vector is $\begin{pmatrix} 2 & -0.5 & 1 \end{pmatrix}'$. The error term u is drawn from a standard normal distribution. Hence theory predicts that the optimal model is the probit model.

Setup 1 ii) uses the following DGP. The latent index is given by $y^* = \sum_{j=1}^{3} \beta_j x_j + u$, where the true β coefficient vector is $(\ 2 \quad -0.5 \quad 1 \ )'$. The error term u is drawn from a logistic distribution, with location parameter 0 and scale parameter 1.[30] Theory predicts that

[30]Some of the (pseudo) random number generators were not implemented in Matlab. If this was the case, I

the logit model is optimal.

Setup 1 iii) has the following DGP. The latent index is given by $y^* = \sum_{j=1}^{3} \beta_j x_j + u$, where the true β coefficient vector is $(\ 2 \quad -0.5 \quad 1 \ )'$. The error term u is drawn from a Cauchy distribution. Hence the first moment of the error term does not exist. This setup is therefore concerned with the performance of the estimators, when the density of the error term has very fat tails.

Setup 1 iv) has the following DGP. The latent index is given by $y^* = \sum_{j=1}^{3} \beta_j x_j + u$, where the true β coefficient vector is $(\ 2 \quad -0.5 \quad 1 \ )'$. The error term u is drawn from a Gumbel distribution. The PDF of the Gumbel distribution is given by the following formula $f(z) = e^{-e^{-(z-\mu)/\beta}}$. The following parameter values were chosen: $\beta = 1$ and $\mu = -0.5772$. The parameters are chosen such that $E(u) = 0$. Due to the fact that the Gumbel distribution has a skewness of ≈ 1.14, this setup is concerned with the performance of the estimators when the error terms are skewed.

Setup 1 v) uses a DGP for the latent index, which is given by $y^* = \sum_{j=1}^{3} \beta_j x_j + u$, where the true β coefficient vector is $(\ 2 \quad -0.5 \quad 1 \ )'$. The error term u is drawn from a symmetric mixture normal distribution. The mixture distribution of u has the PDF $f(z) = \frac{1}{2}\phi(z+2) + \frac{1}{2}\phi(z-2)$, where $\phi(\cdot)$ denotes the PDF of a standard normal distribution. Hence the mixture distribution used consists of two equally weighted normal distributions with mean -2 and 2, and variance of 1. The PDF of the distribution is depicted in Figure 2. Since the distribution is bimodal[31] the setup is concerned with errors that are bimodal which results in a non-standard link function.

Setup 1 vi) uses a DGP, which is given by $y^* = \sum_{j=1}^{6} \beta_j x_j + u$, where the true β coefficient vector is $(\ 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \ )'$. The error term u is drawn from a standard normal distribution. Hence theory again suggests that the optimal model is the probit model. First, this setup was constructed to check the robustness of the results for more than three non-constant regressors. Second, this setup will shed some light on the differences in the performance of the semiparametric estimators, when the dimension of X changes. From a theoretical perspective this change should not be substantial. This stems from the fact that the estimators reduce the dimensionality of their nonparametric part through the single index stucture.[32]

Setup 1 vii) uses the following DGP, which is given by $y^* = \sum_{j=1}^{3} \beta_j x_j + u$, where the true β coefficient vector is $(\ 2 \quad -0.5 \quad 1 \ )'$. The error term consists of a mixture of three normal distributions. 10% of the data are generated by a normal distribution with mean -1

programmed the random number generators using the uniform random number generator and evaluating the inverse of the distribution function of interest. The method is described in the appendix.

[31] The modes are near -2 and 2.

[32] The semiparametric estimators use $X_i'\theta$ which is a scalar instead of X_i which is k-dimensional for the nonparametric estimation.

and standard deviation 1, 80% stem from a standard normal distribution and the remaining 10% come from a normal distribution with mean 1 and standard deviation 1. Hence the setup is not concerned with single classical isolated outliers. However, it is likely by construction that some observations will occur as if they were pure outliers, especially in smaller samples.

Setup 2 i) uses $y^* = \sum_{j=1}^{2} \beta_j x_j + \beta_3 ln(x_3) + u$ as the DGP. The error terms were drawn from a standard normal distribution. Since the natural logarithm $ln(\cdot)$ is not defined for negative inputs, the regressors x_3 are drawn from the lognormal distribution, whose support is positive. To make the setup similar to the first seven setups I transformed the random variables and the coefficients such that $E(y^*) = 0$ and hence $E(p) = 0.5$.[33] In the estimation, the variable x_3 instead of $ln(x_3)$ enters the index. This setup is motivated through the fact, that in empirical applications one usually does not know the exact relation between x and y. Since the functional form of the index is misspecified, it is unlikely that any of the estimators will estimate the marginal effects consistently. However, since the $ln(\cdot)$ is a positive monotone transformation, the models should be able to estimate the sign of the marginal effects consistently.

Setup 2 ii) has the following DGP. $y^* = \left(\sum_{j=1}^{3} \beta_j x_j\right)^3 + u$. u is drawn form a standard normal distribution and $\beta = (\ 2 \quad -0.5 \quad 1\ )'$. Setup 2 ii) is similar to setup 2 i). Again we assume that the exact form of the index function is not known. Hence in the estimation $\sum_{j=1}^{3} \beta_j x_j$ is used as the index. The interest focuses on the question if the sign of the marginal effects is estimated consistently given positive monotone transformations of the index.

Setup 2 iii) has the DGP $y^* = \sum_{j=1}^{4} \beta_j x_j + u$. The first two regressors and u are drawn from standard normal distributions. $\{x_{i3}, x_{i4}\}_{i=1}^{N}$ are drawn from a multivariate normal distribution with correlation $\rho = 0.5$. The coefficient vector is $\beta = (\ -1 \quad -.5 \quad 1 \quad -2\ )'$. For the estimation it is further assumed that $\{x_{i4}\}_{i=1}^{N}$ is unknown. In a linear model, the omitted variable would lead to a downward bias in $\hat{\beta}_3$. Since the link function is monotone, I expect that the marginal effects are as well downward biased.

[33] This was done as follows. x_1 was drawn from a normal distribution with expected value of one, x_2 was drawn from a standard normal. x_3 was drawn such that $ln(x_3)$ had an expected value of one. Since $\beta = (\ -1 \quad -0.5 \quad 1\ )'$, the expected value of y^* is zero.

4 Results

This chapter describes the results of the ten different Monte Carlo setups. Since the bandwidth choice and the calculation method of the marginal effects for the LLL estimator have a huge effect on the estimates, the chapter starts with a discussion of the specifications of the LLL estimator. Afterwards, the discussion of the results of the ten different Monte Carlo setups focus on three aspects. First, the predictive performance of $\hat{y}$ is discussed. Second, the results for the average marginal effect, the marginal effect at the average and the marginal effects at the first and third quartile are presented. Finally, for the Monte Carlo setups S1 i) - S1) vii), the distributions of the average marginal effects are presented. The reason why this presentation seems relevant is the fact that theory predicts that some estimators should have asympotically normally distributed average marginal effects.[34] Since it is not a priori obvious which sample size is sufficient, such that the estimators of the marginal effects are more or less normal distributed, the visual presentation might help to get an idea, when this is the case.[35] The asymptotic normal distribution of the average marginal effect estimators is at least in two ways useful. First, it allows to conduct usual t-tests. Second, a test proposed in chapter 5 which helps deciding between a parametric and a semiparametric estimator is based on the assumption of normally distributed estimators of the average marginal effects.

The amount of results produced makes it impossible to describe them completely within the thesis. Since a graphical presentation has the advantage to describe a great number of results in an accessible manner, I decided to present the results via graphs. As the disadvantage of the graphical presentation is less accuracy I offer to give the exact results on request. Further, I made the decision to present the graphs within the text. This shall enable the reader to form his own opinion regarding the results of each setup directly (without switching from the text to the appendix). The results of the first Monte Carlo setup presented in chapter 4.2. will be described extensively, while the remaining setups will be described more briefly. Still the graphical presentation allows the reader to form his own opinion of the results of each setup. For the reader exclusively interested in the main results I suggest to skip the description of the single setups and to go directly to the end of chapter 4.8. The main conclusions regarding the results of the first seven setups are given there. For the reader specifically interested in a particular setup I suggest to read chapter 4.2. first and then to go to the setup of interest.

[34] Given that the coefficient estimates are asymptotically normal, the parametric estimators should deliver estimates of the average marginal effects which are asymptotically normally distributed. Further, Klein and Spadys estimator should as well have asymptotically normal estimated average marginal effects. A heuristic explanation for this is given in the appendix.

[35] It should be kept in mind that due to the full dependence of the results on the assumed DGP extrapolation of the results might not be valid.

4.1 Bandwidth choice and marginal effects method for the LLL estimator

In the following, I will present the results of six different specifications for the LLL estimator. The results presented are the RMSE, the average marginal effect and the marginal effect at the first quartile for the standard setup S1 i) where errors are normally distributed. The specifications are three choices of the bandwidth combined with two ways of calculating the marginal effects. The bandwidth choice was half ($h = 0.5S$), twice ($h = 2S$) Silverman´s plug-in estimate and Silverman´s plug-in estimate itself ($h = S$) as discussed in chapter 2.2.1. The calculation of the marginal effects was performed either over finite differences (FD) or the partial derivative (PD) as discussed in section 2.2.3.

Table 10: Results of the LLL specifications

Quantity	Specification	n=100	n=500	n=1000
RMSE	$h = S$	0.22	0.27	0.28
	$h = 0.5S$	0.06	0.15	0.18
	$h = 2S$	0.28	0.29	0.3
AvME	$h = S$, PD	-8.49	0.90	0.49
	$h = S$, FD	0.13	0.16	0.17
	$h = 0.5S$, PD	-183.81	0.08	0.33
	$h = 0.5S$, FD	0.04	0.09	0.14
	$h = 2S$, PD	0.33	0.41	0.38
	$h = 2S$, FD	0.17	0.16	0.16
	True	0.16	0.16	0.16
MEQ1	$h = S$, PD	0.16	0.05	0.04
	$h = S$, FD	1.82	0.10	0.07
	$h = 0.5S$, PD	-0.0003	-0.077	0.2813
	$h = 0.5S$, FD	0.09	-0.25	0.003
	$h = 2S$, PD	0.063	0.077	0.074
	$h = 2S$, FD	0.069	0.084	0.083
	True	0.10	0.10	0.10

As one can see from Table 10, the estimated marginal effects are closest to the true value, when the bandwidth choice is the largest and the finite difference method is used. Further, the results with respect to the RMSE are best when the bandwidth is small. At first glance this might seem surprising. Hence, I try to give an intuitive explanation for the results, but do not attempt to formalize the ideas. As the bandwidth choice decreases the local character of the estimation becomes more pronounced. In the limit, as $h \to 0$ the local estimator takes only one value of X into account. If the value for X happens to be in the sample, the local estimator assigns the corresponding value of y for $\hat{y}$,[36] if it is not, the corresponding value is zero. Mechanically, this results in an estimation where $\hat{y} \to y$, hence the following assertion

[36]More exactly the $\hat{\theta}$´s are chosen such that $\hat{y}$ takes the value of y. As an aside there will be an identification problem for $h \to 0$, since only one sample point is used to estimate k parameters. Additionally this only holds if the values X is taken by only one sample point. However the idea stays the same if there are several sample points with the same value of X.

seems reasonable. As h tends to zero, the RMSE should tend to zero as well. On the other hand as h becomes small the changes in the prediction of y are rapid. Again looking at the limiting case might help to get an intuition. As one changes X marginally, starting from a point in the sample with $y = 1$, it is most likely that for $h \approx 0$ the predicted value jumps from 1 to 0. Therefore, the estimation of the marginal effects via finite differences does not seem accurate for small h. Again, as discussed in section 2.2.3, when considering a uniform kernel and $h \approx 0$ only few values in the sample are taken into account to estimate k parameters of the vector θ. Therefore, it is no surprise that the estimation of θ becomes inaccurate. This reasoning might explain the poor performance of the analytical derivative for small h. As a result the LLL estimator is considered with bandwidth equalling twice Silverman´s plug-in estimate and the finite difference method is used for the estimation of the marginal effects.

4.2 Setup 1 i): Normally distributed errors

Figure 6 describes the performance of the estimators with respect to their predicted values $\hat{y}$. The upper left graph describes the mean of the RMSE for the five estimators given in the legend as the sample size varies. The upper right graph describes the evolution of the mean of the RMSE for the inner 80% of the sample. The lower left graph displays the mean of the RMSE using zero-one predictions. The lower right graph describes the share of "wrong" predictions given that $\hat{p}_i > 0.95$. The results for the OLS estimator (used in the linear probability model) are not displayed. The RMSE´s of OLS take on values between 0.53 and 0.63 and given that $\hat{p}_i > 0.95$ no individual had a value of $y_i = 0$. The main results of the graphs are the following. Regarding the RMSE, the LLL estimator outperforms all other estimators. This is not surprising given the local character of the estimation. This result is very robust across setups and will therefore not be discussed in the remaining setups. The probit and logit estimators have nearly the same RMSE. The cauchit estimator becomes worse and Klein and Spady´s estimator becomes relatively better as the sample size increases. The results for the inner sample (RMSE80, upper right corner) are similar. Given zero-one predictions (lower left corner) cauchit, KS and LLL outperform the logit and probit estimators. The lower right graph depicts the share of "wrong" predictions, where "wrong" means that an individual with $\hat{p}_i > 0.95$ has a value of $y_i = 0$. This measure should be smaller than 5%. All estimators fulfill this criterion, they all lie below 1.2%.

Figure 6: Performance of $\hat{y}$ given u is normal

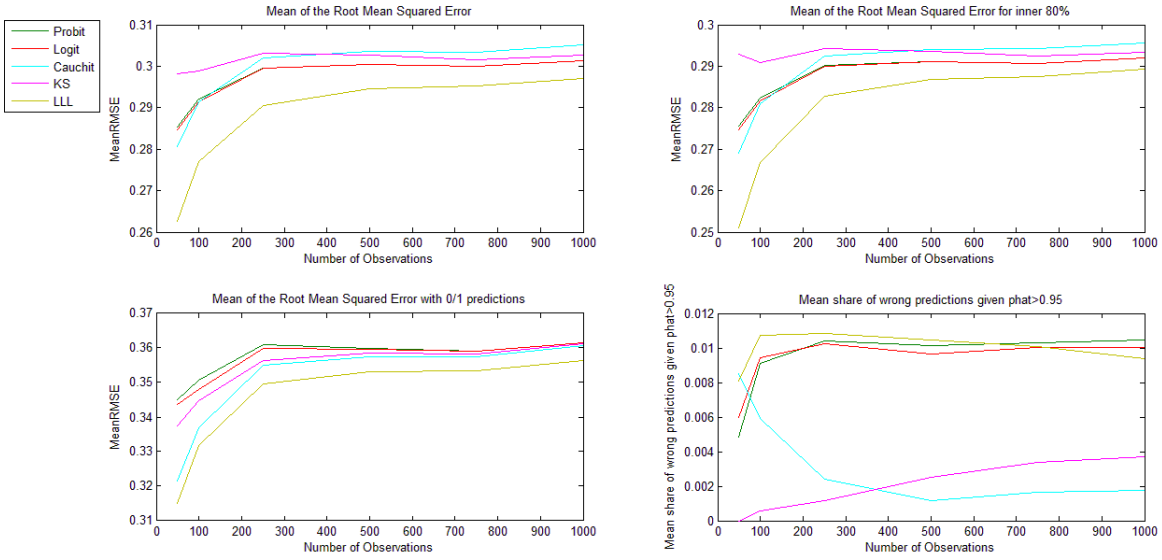


Figure 7 depicts the estimates of the average marginal effect and the marginal effect at the average. The upper graphs display the means, whereas the lower graphs display the standard deviations. All estimators, except the cauchit estimator, appear consistent for the **average marginal effect**. Klein and Spady´s estimator improves substantially when the sample size increases, however the sample size was insufficient to assess formal consistency. The probit and

logit model have a low variance. As theory would suggest, it seems that the probit estimator is asymptotically efficient.

The only estimator which appears consistent for the **marginal effect at the average** is the probit estimator. The deviations of the estimates from the LLL, KS and logit estimator to the true value are minor. KS´s estimator improves substantially when the sample size increases. Due to the fact that the sample size was not sufficiently large a final assessment of KS´s consistency is not possible. Since the deviations of the estimates from OLS and the cauchit model to the true value are large, their performance with respect to the marginal effect at the average is poor.

Figure 7: Average ME and ME at average for u normal

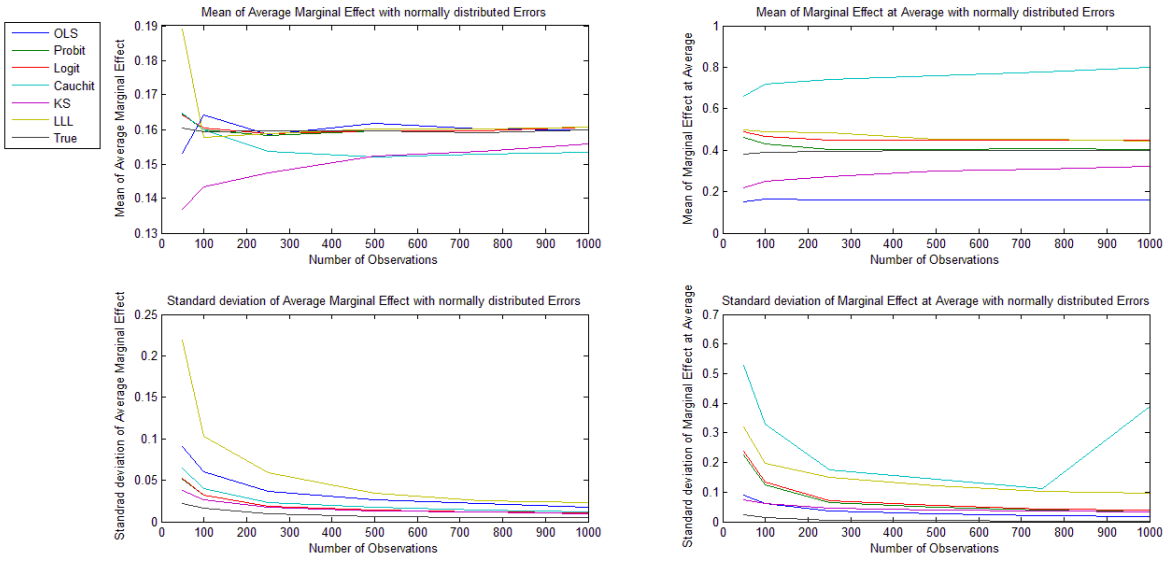


Figure 8 displays the mean and the standard deviation of the estimates of the marginal effects at the quartiles. As for the marginal effect at the average, the only consistent estimator for the **marginal effects at the quartiles** is the probit estimator. Again the sample size is insufficient for the assessment of KS´s consistency. The deviations for KS, LLL and logit to the true value are minor, whereas OLS and cauchit perform poorly. Moreover it is worth noticing that the standard deviation of the estimated marginal effect at the quartiles from the LLL model does not decrease uniformly with growing sample size.

As one will see later on, the pattern above repeats over most of the setups in the Monte Carlo study. Most estimators perform well in estimating the average marginal effect and in general the performance of the parametric estimators, except the one with properly specified likelihood, is poor for the marginal effect at the mean or the quartiles.

Figure 8: ME at first quartlile and third quartile for u normal

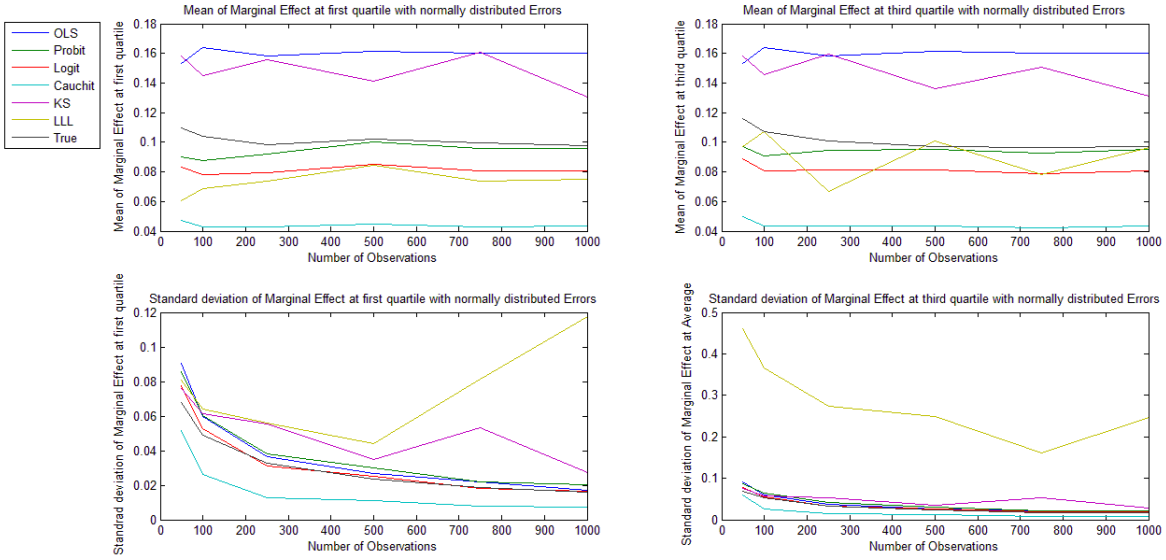


Figure 9 to 11 depict the distribution of the estimates for the average marginal effects. The two axis at the bottom describe the sample size and the value of the average marginal effect. The vertical axis describes the value of a density estimate. Hence, slices along the axis, holding the number of observations constant, describe an estimate of the density of the average marginal effect estimator. The upper graphs attempt to present the distributions for small samples and the lower graphs focus on the distribution for large samples. The discussion is very informal. The advantage of the graphs is that the behaviour of the estimators for the average marginal effect, for different sample sizes, is easily seen. A disadvantage is that the tail behaviour is hard to interpret. Therefore it might be the case that an estimator appears normally distributed in the picture, but has substantial excess probability mass in the tails. This limitation should be kept in mind.

The main points to notice are the following. The distribution of the estimated average marginal effects of the parametric estimators appears normal, even for relatively small sample size. Klein and Spady´s estimator produces estimates of the average marginal effect which seem non-normal for small samples but become normally distributed as the sample size increases. Even in large samples the LLL estimator appears non-normal at the tail of the distribution (around 0.2). The relative frequency does not decline smoothly. However, with the exception of the tail behaviour the distibution appears normal. This result could be due to the limited number of repetitions (R=200) in the Monte Carlo study.

Figure 9: Distribution average ME, Probit and Logit, u is normal

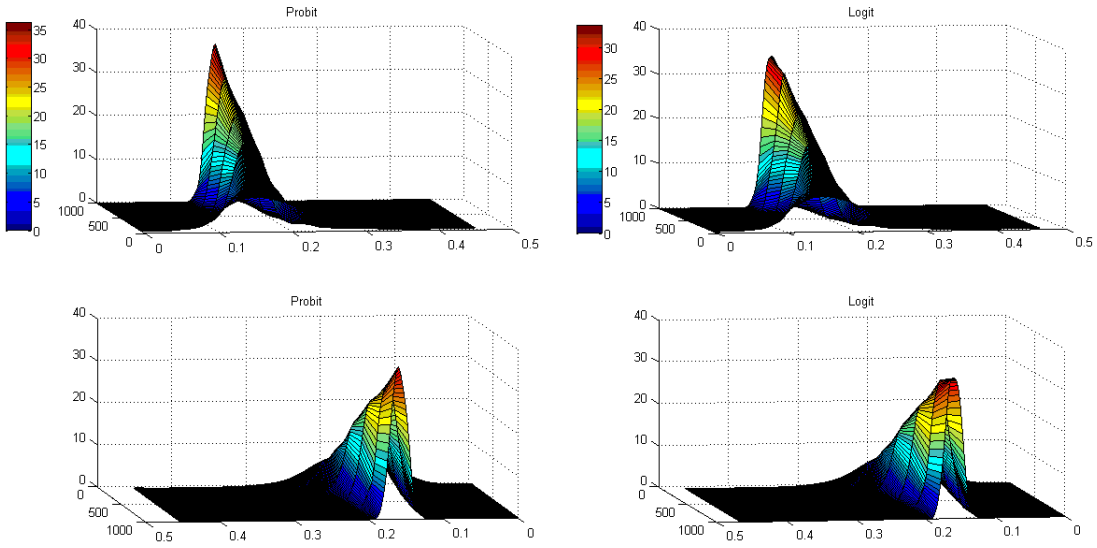


Figure 10: Distribution average ME, OLS and Cauchit, u is normal

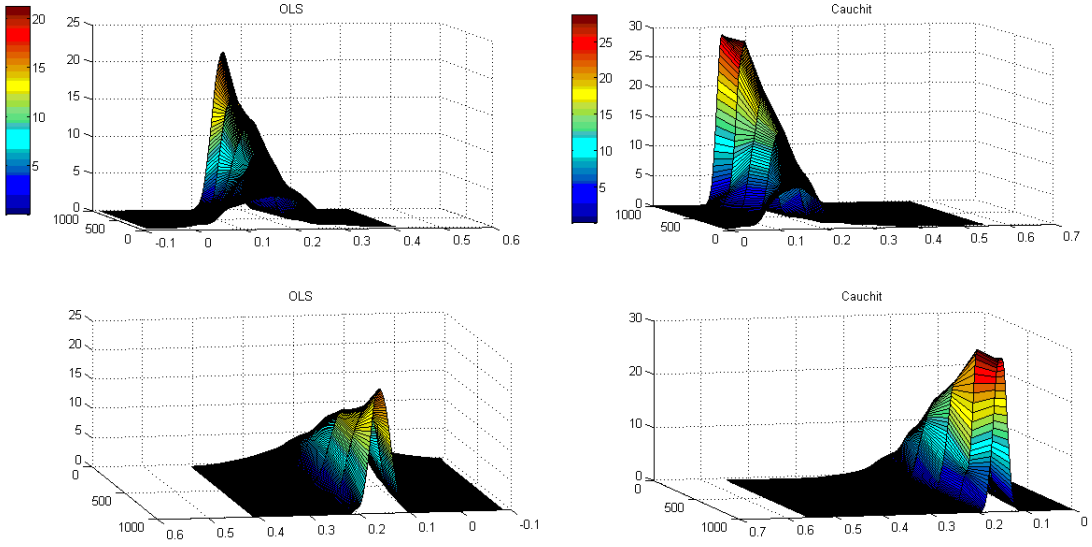
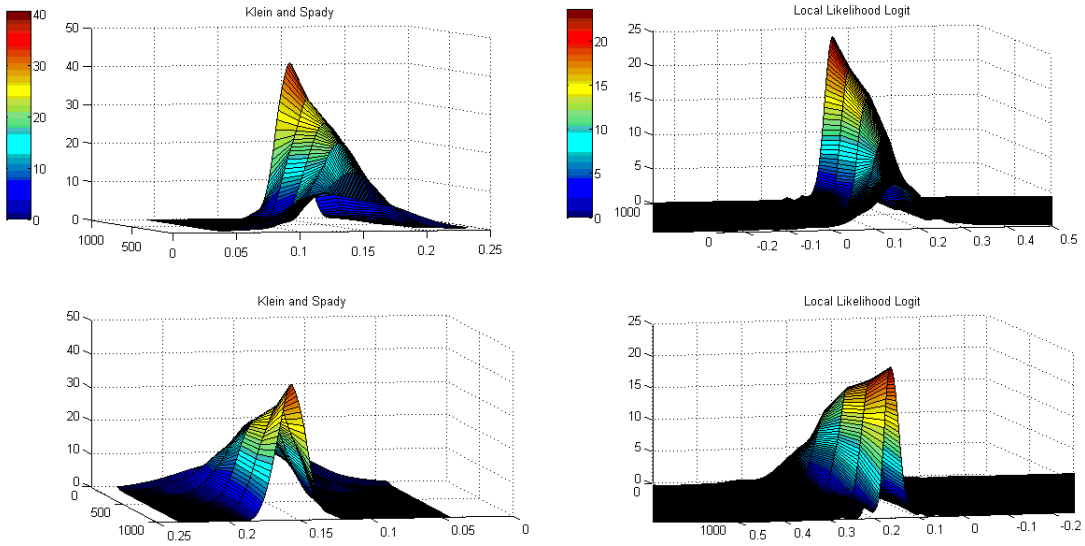


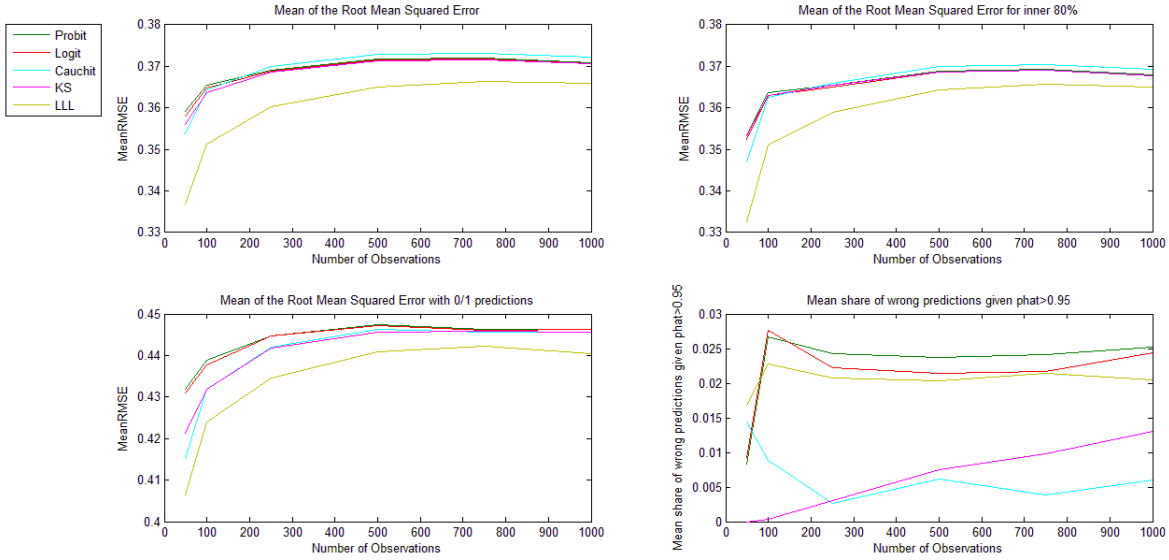
Figure 11: Distribution average ME, KS and LLL, u is normal



4.3 Setup 1 ii): Logistic distributed errors

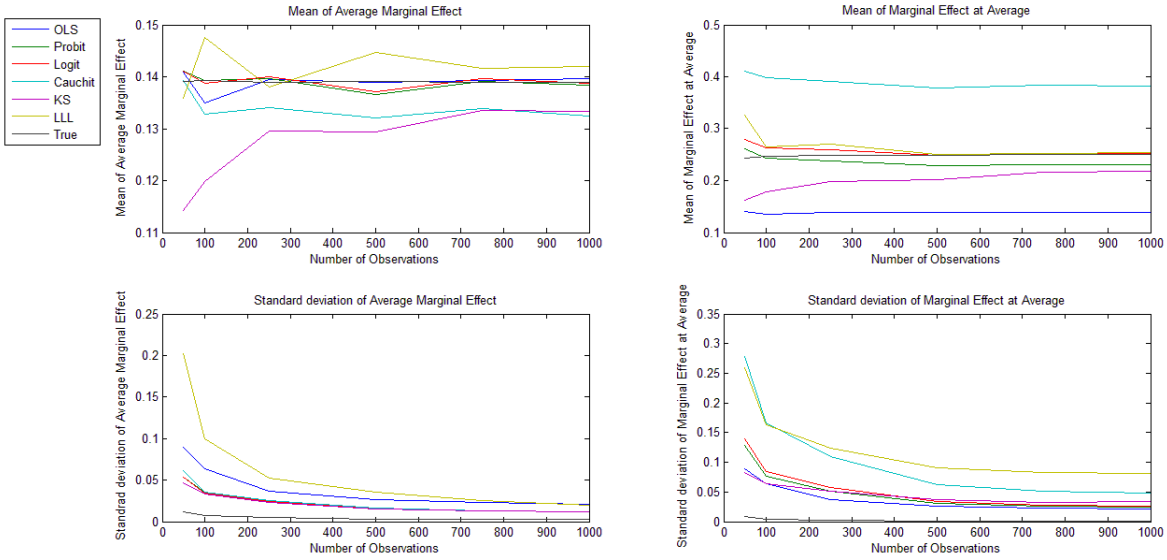
The results for the estimators of $\hat{y}$ are similar to those of the previous setup. The LLL estimator performs best. The estimators derived from the logit and probit model have nearly the same RMSE. Looking at zero-one predictions the cauchit-, and Klein and Spady´s estimator outperform the logit and probit estimators. The mean share of wrong predictions is lower than 3% for all estimators. Note though that the mean share of wrong predictions is higher than in the previous setup in general. The OLS estimator (not displayed) has a RMSE between 0.55 and 0.65 and the share of wrong predictions is zero in all samples.

Figure 12: Performance for $\hat{y}$ given u is logistic



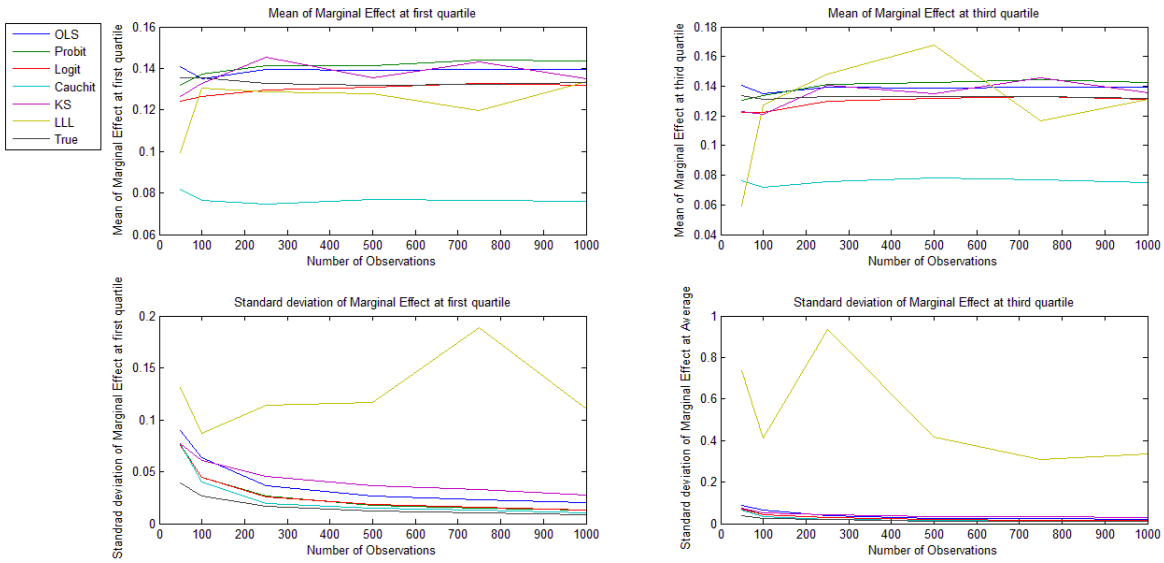
As in the previous setup all the estimators perform well with respect to the average marginal effect. Only Klein and Spady´s estimator deviates substantially from the true value in small samples, but improves as the sample size increases. OLS and cauchit substantially underestimate or overestimate the marginal effect at the average. The LLL estimator has the highest variability.

Figure 13: Average ME and ME at average for u logistic



As expected, the logit estimator estimates the effects at the quartiles consistently. The deviations of the remaining estimators, except cauchit, are of magnitude smaller than 0.02. At least for the OLS estimator this result seems to be a coincidence, resulting from the fact that the true average marginal effect nearly coincides with the marginal effects at the first and third quartile. As for the setup with normally distributed errors, the standard deviation of the LLL estimator seems not to diminish uniformly as the sample size grows.

Figure 14: ME at first quartlile and third quartile for u logistic



The distributional pattern of the estimated average marginal effects is mainly similar to those resulting from the setup with normally distributed errors. All estimators appear normally distributed. Moreover the shapes of KS´s and the LLL estimator look more like a normal distribution than in the setup with normal distributed errors.

Figure 15: Distribution average ME, Probit and Logit, u is logistic

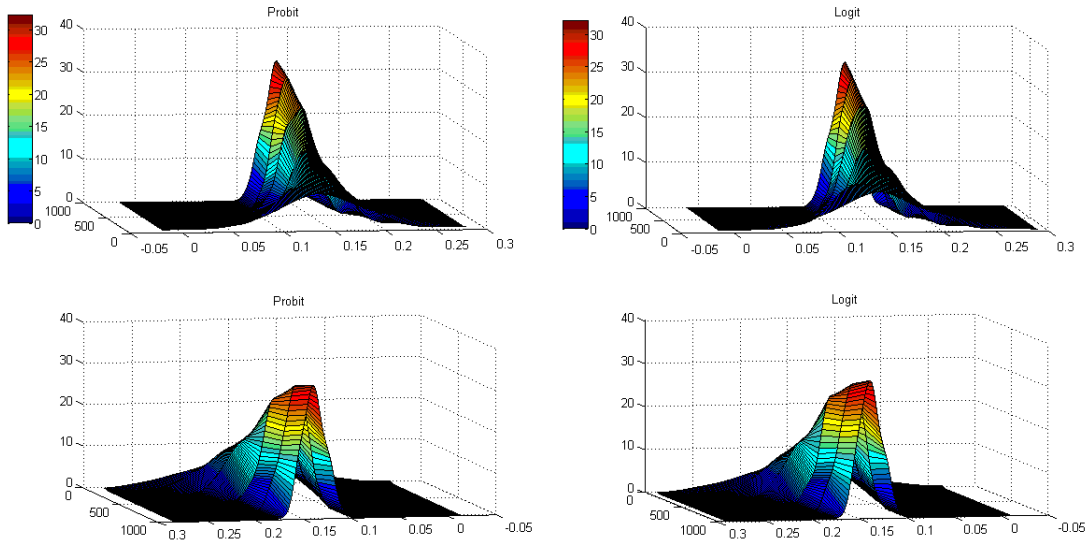


Figure 16: Distribution average ME, OLS and Cauchit, u is logistic

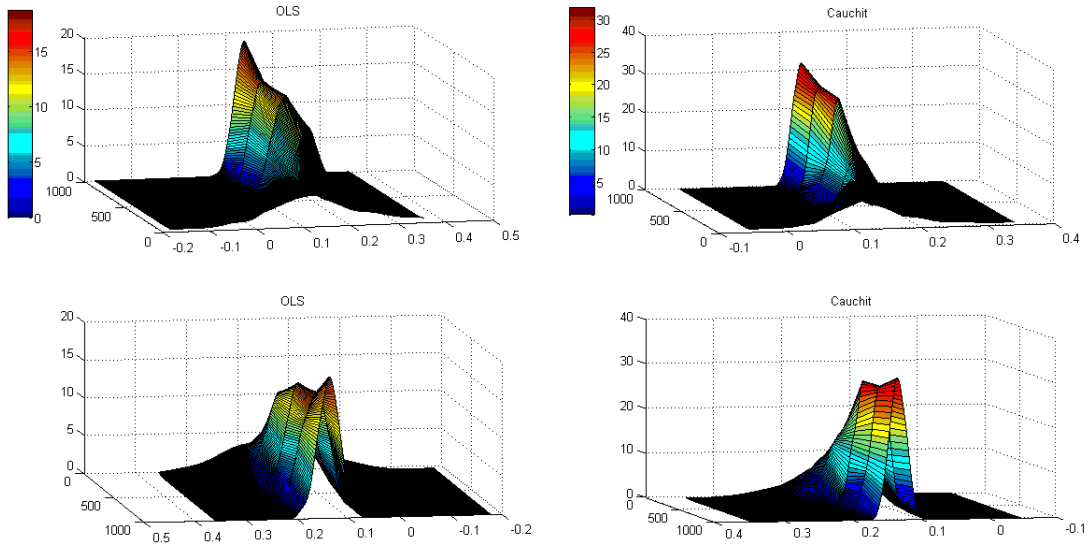
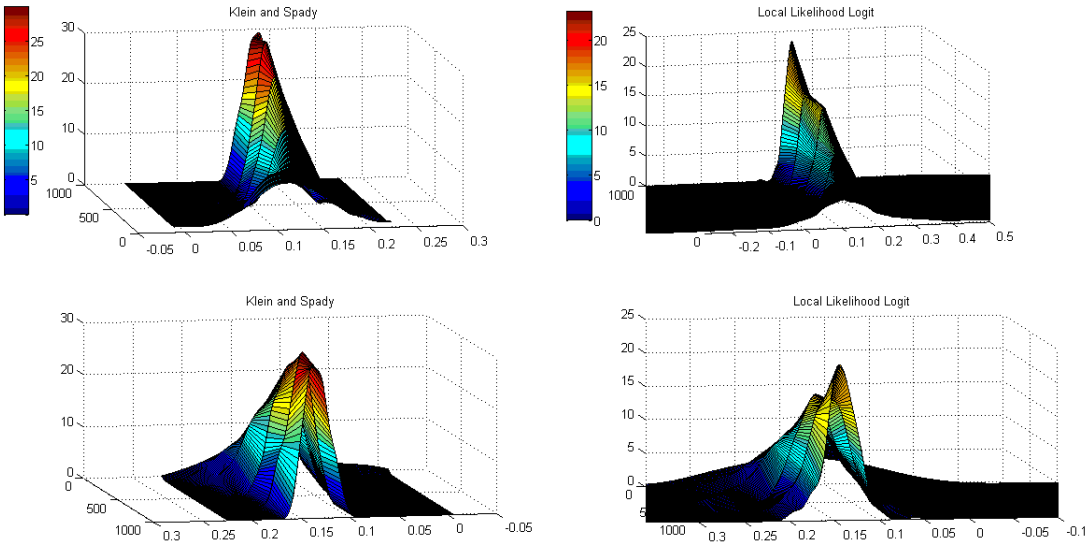


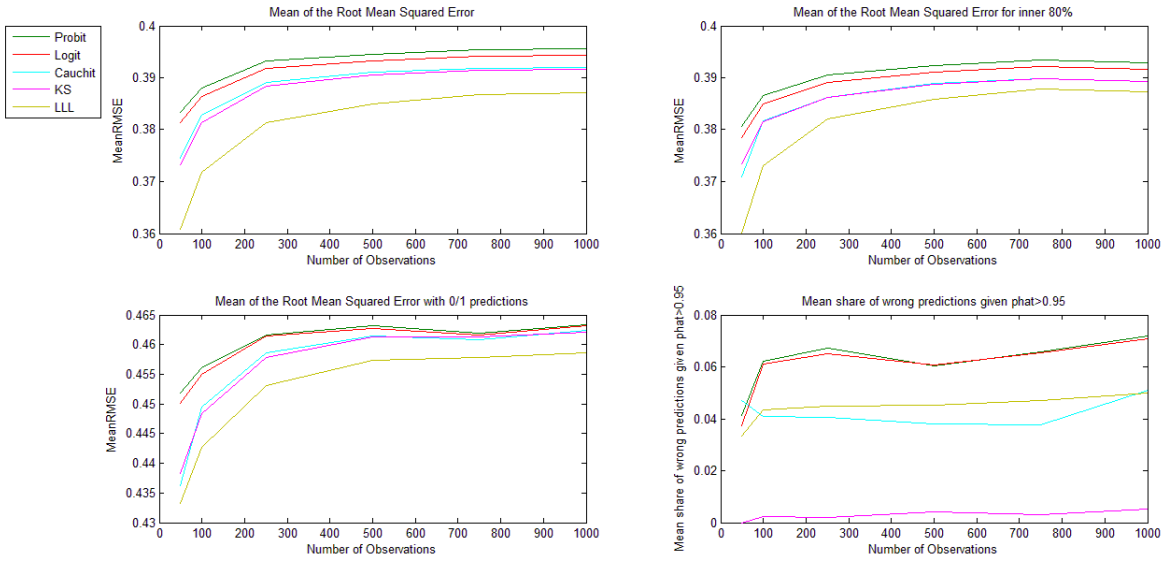
Figure 17: Distribution average ME, KS and LLL, u is logistic



4.4 Setup 1 iii): Cauchy distributed errors

Running the simulation with fat tailed, Cauchy distributed errors, the semiparametric estimators outpeform the parametric ones with respect to the RMSE. In the class of parametric estimators, the theoretically efficient cauchit estimator outperforms both logit and probit. The mean share of "wrong" predictions varies substantially across estimators. KS´s share of wrong predictions is below 1%, Cauchit´s and LLL´s near 5 % and the share of wrong predictions for the logit and probit model is generally above 5%. The OLS estimator (not displayed) has a RMSE of around 0.55 to 0.65 and a share of wrong predictions substantially above zero, more exactly the share increases from 1% to 4% as the sample grows. From setup 1 i) - setup 1 iii) a potential hypothesis could be that higher kurtosis of the error term leads to a higher share of "wrong" predictions in the upper tail of $\hat{y}$. However, as we will see in the next setup the share of "wrong" predictions at the upper tail is lower for Gumbel distributed errors than for logistic distributed errors (even though the Gumbel distribution has a higher kurtosis). Thus, the effect of the kurtosis of the error term on the share of "wrong" predictions seems unclear.

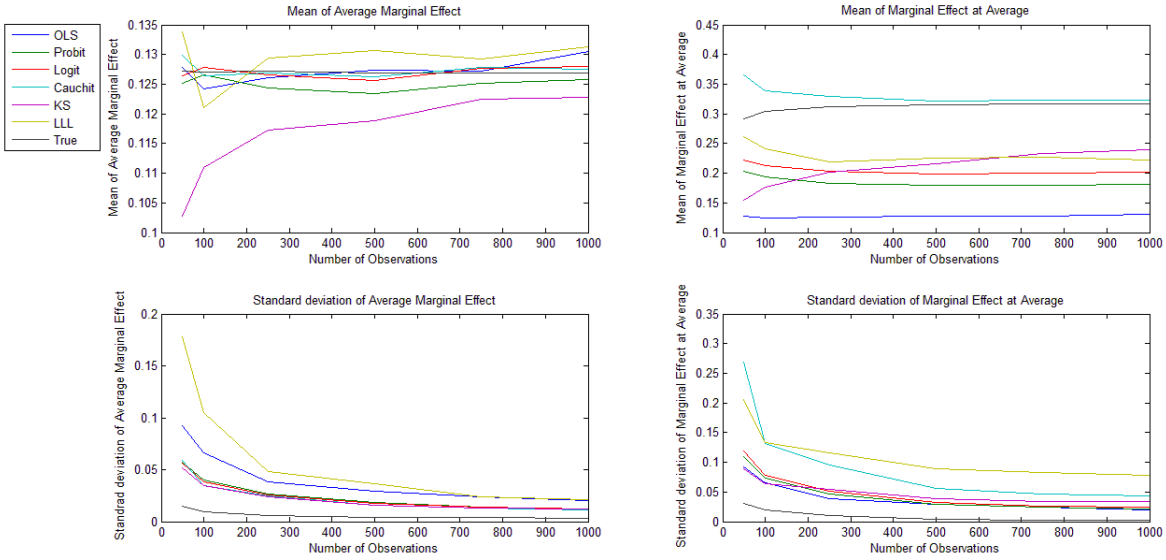
Figure 18: Performance for $\hat{y}$ given u is Cauchy



Surprisingly for the first non-standard distribution of the error term, the picture for the average marginal effect looks similar to those presented before. Speaking from consistency in a strict statistical sense is problematic. However the deviations of the estimators from the true model are minor. Initially KS´s estimator performs poorly, but as before becomes substantially better with growing sample size.

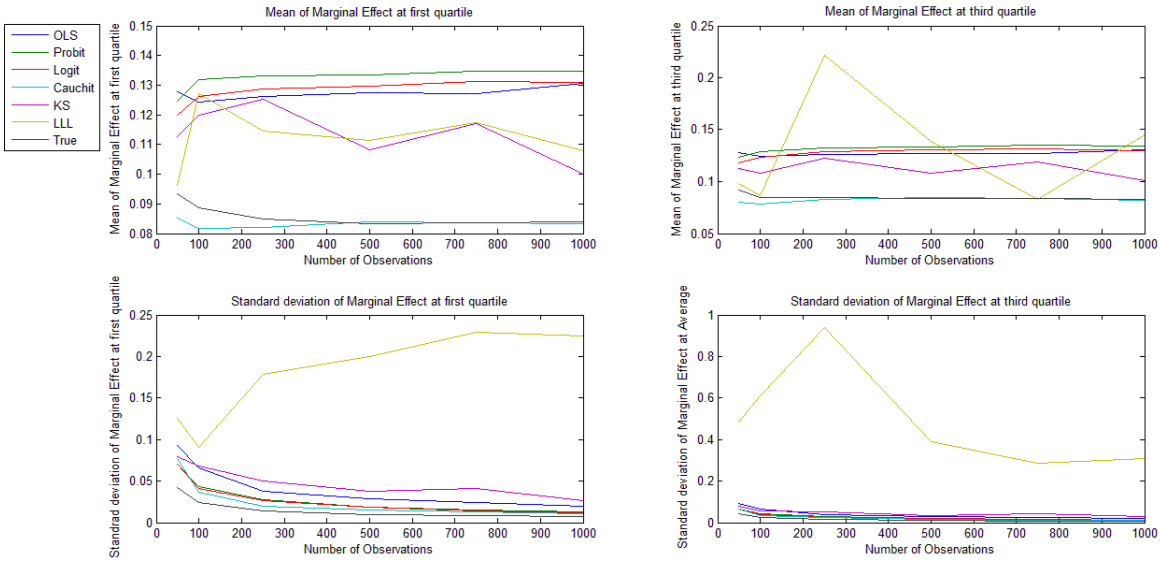
Looking at the estimated marginal effects at the mean reveals, that the cauchit estimator performs best, whereas for large N the two semiparametric estimators are ranked second and third. The remaining parametric estimators perform poorly.

Figure 19: Average ME and ME at average for u Cauchy



As above the best performing estimators for the marginal effects at the first and third quartile are, cauchit and KS. Since the standard deviation of the LLL estimator is not uniformly decreasing, the hypothesis of LLL being consistent for the estimation of the marginal effects at the quartiles is not supported.

Figure 20: ME at first quartlile and third quartile for u Cauchy



The results for the distribution of the estimated average marginal effects displayed in Figure 21-23 are the following. As for the setups before the shape of the distributions looks for not too small samples similar to a normal distribution.

Figure 21: Distribution average ME, Probit and Logit, u is Cauchy

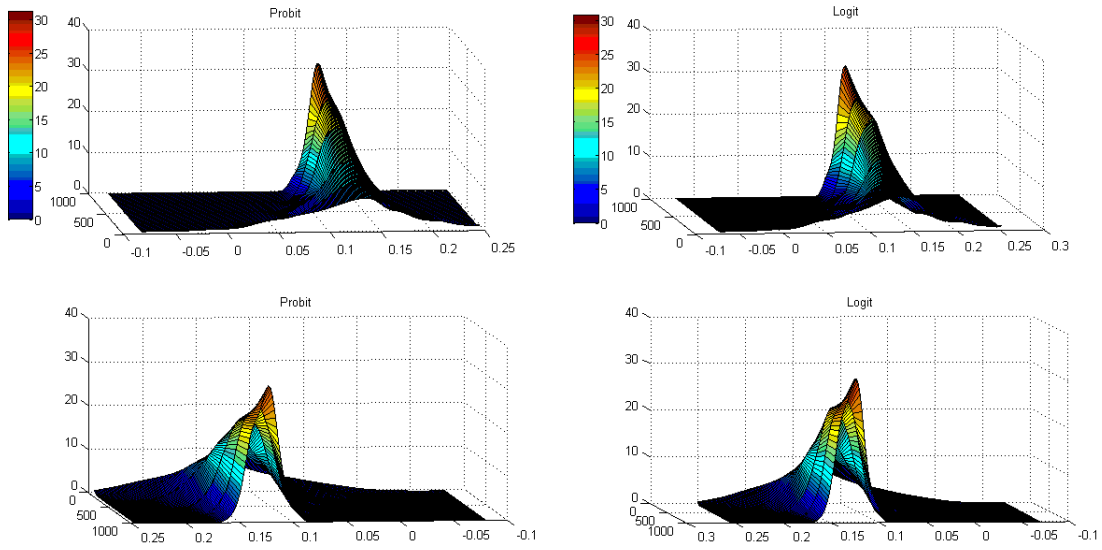


Figure 22: Distribution average ME, OLS and Cauchit, u is Cauchy

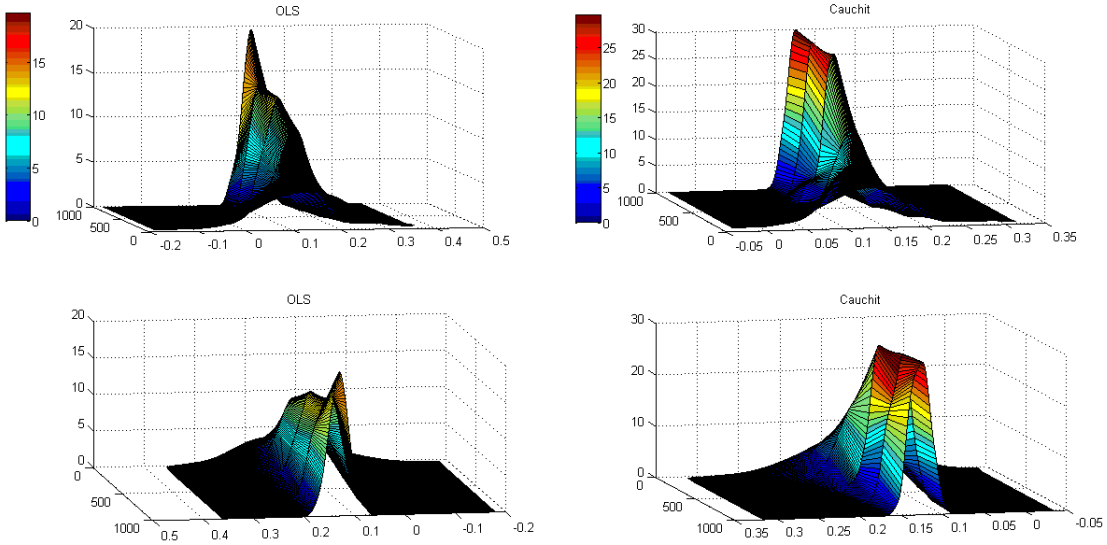
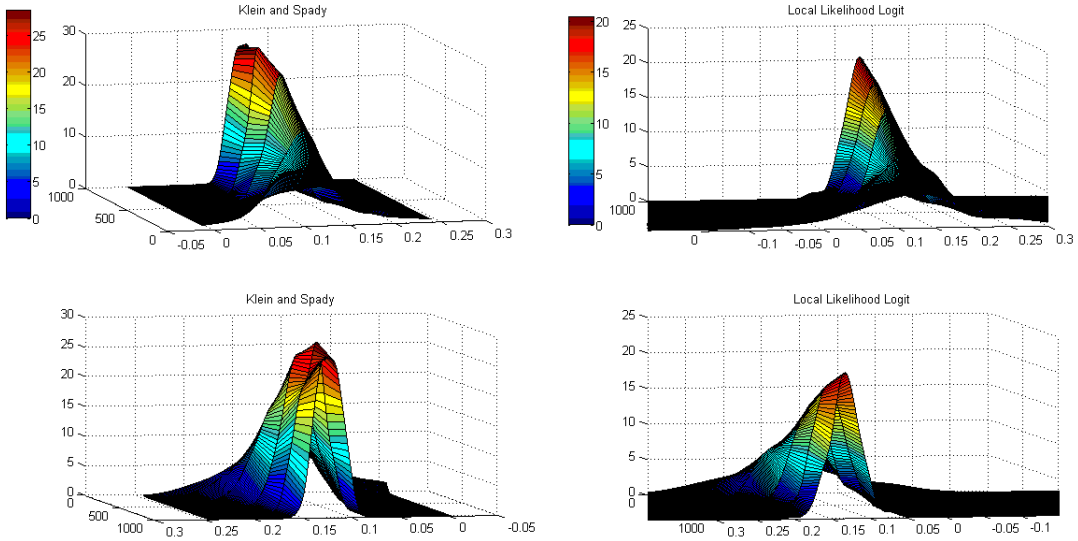


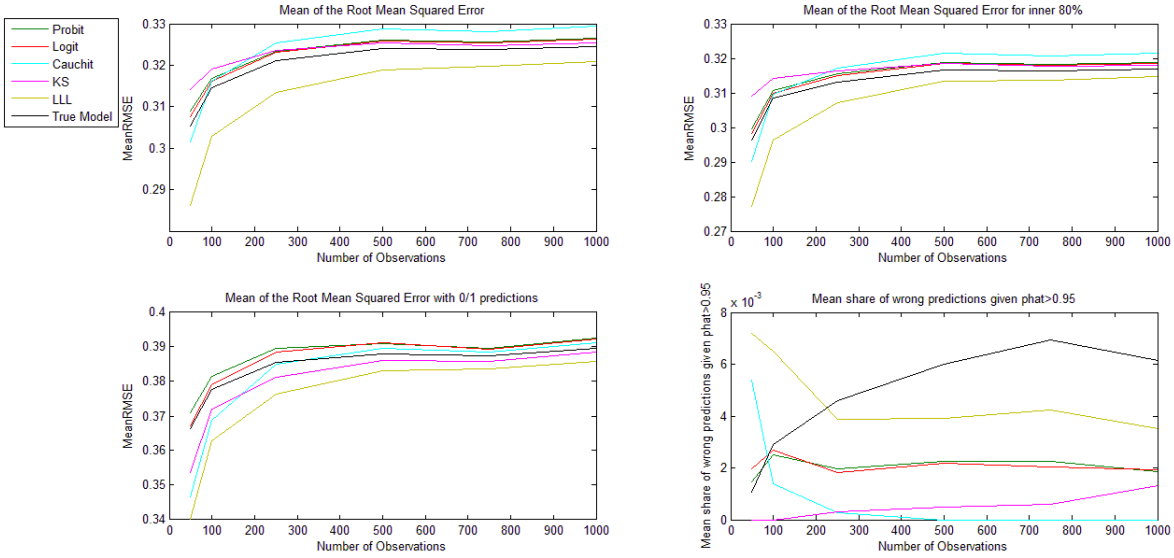
Figure 23: Distribution average ME, KS and LLL, u is Cauchy



4.5 Setup 1 iv): Gumbel distributed errors

Next the results for errors drawn from the skewed Gumbel distribution are presented. The performance ranking for the RMSE (Figure 24) is the following. The LLL estimator performs best. Second best is the "true model" which uses the appropriate transformation of the Gumbel distribution as the link function. Third ranks Klein and Spadys estimator. For Gumbel distributed errors the share of wrong predictions lies consistently below 1%.

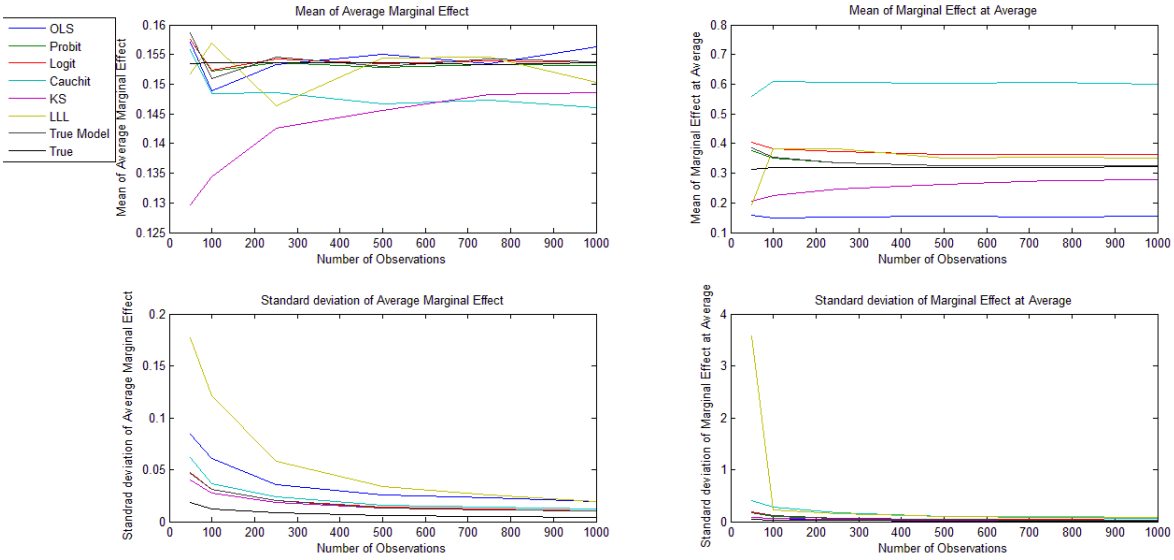
Figure 24: Performance for $\hat{y}$ given u is Gumbel



For the mean of the estimated average marginal effects (Figure 25) the differences across the estimators are minor. The parametric estimators, with the exception of the cauchit estimator, perform very well. As before KS´s estimator performs better, when the sample size increases.

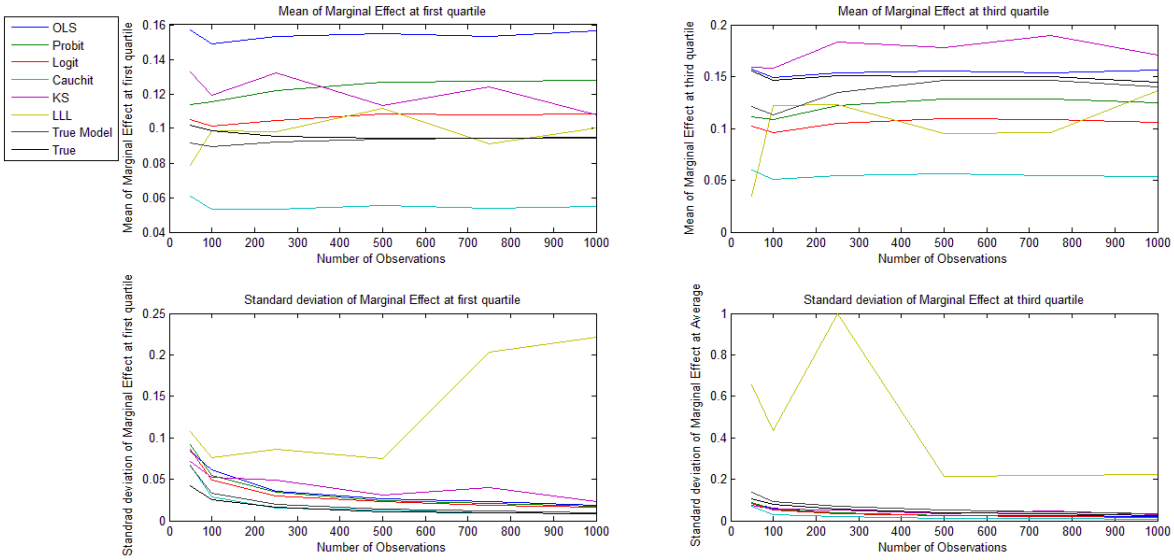
For the marginal effect at the average, the ranking is as follows. The model which uses the true link function (true model) performs best, followed by the probit-, the LLL- and KS´s estimator. The deviations for the cauchit and the OLS estimator are sizeable.

Figure 25: Average ME and ME at average for u Gumbel



The true model performs best, when one considers the marginal effects at the quartiles. Klein and Spady´s estimator again becomes better as sample size increases. The LLL estimators standard deviation does not decrease as the sample size grows. The performance of the cauchit estimator is worst.

Figure 26: ME at first quartlile and third quartile for u Gumbel



Figures 27-30 below show symmetric distributions, which appear normal.

Figure 27: Distribution average ME, Probit and Logit, u is Gumbel

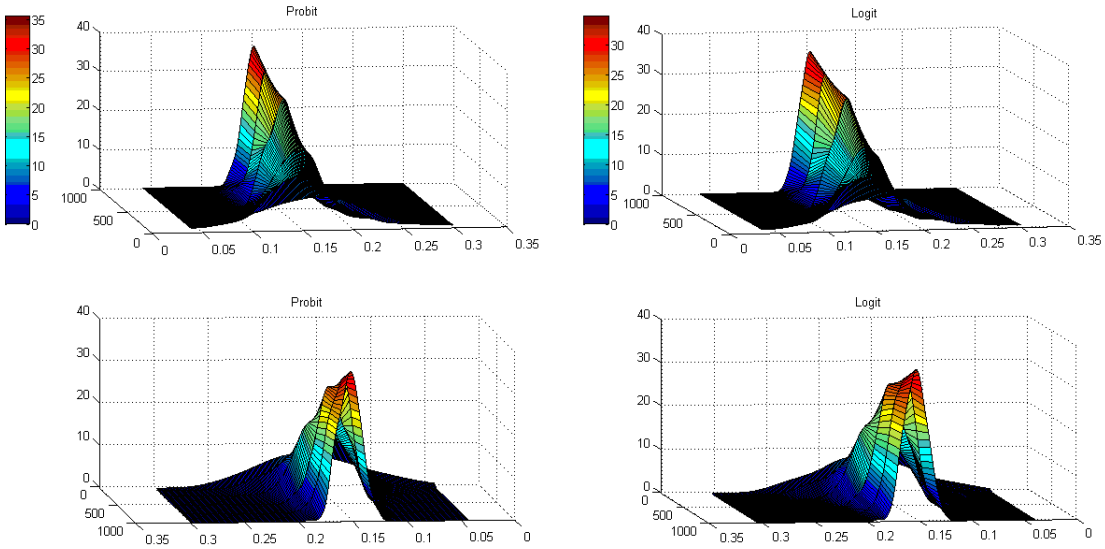


Figure 28: Distribution average ME, OLS and Cauchit, u is Gumbel

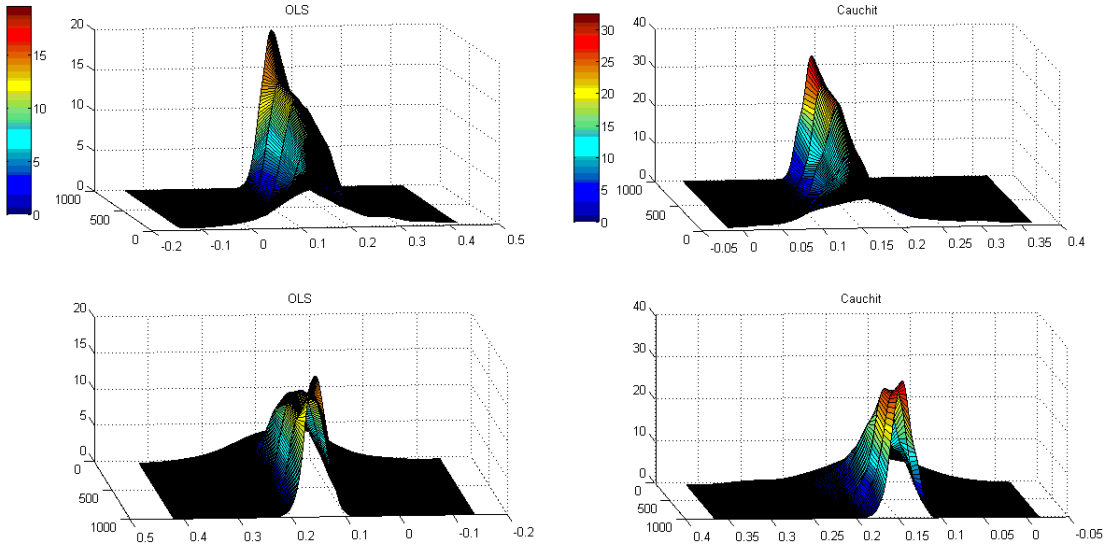


Figure 29: Distribution average ME, KS and LLL, u is Gumbel

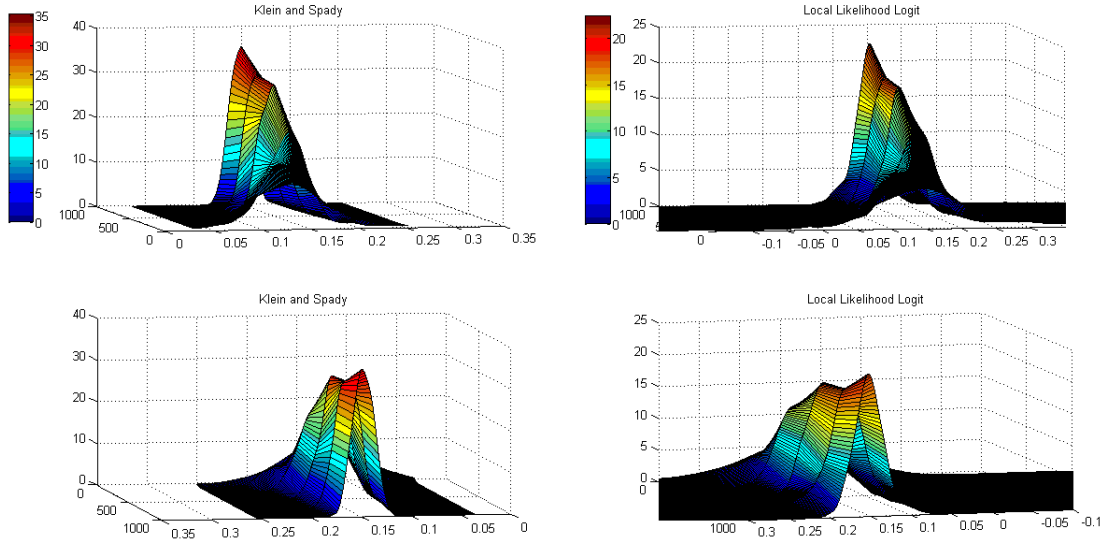
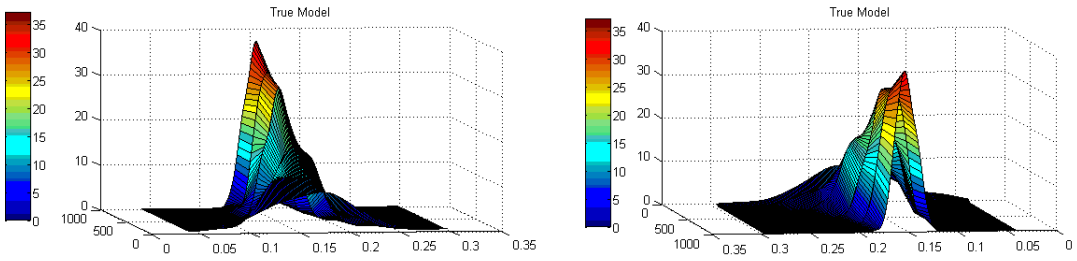


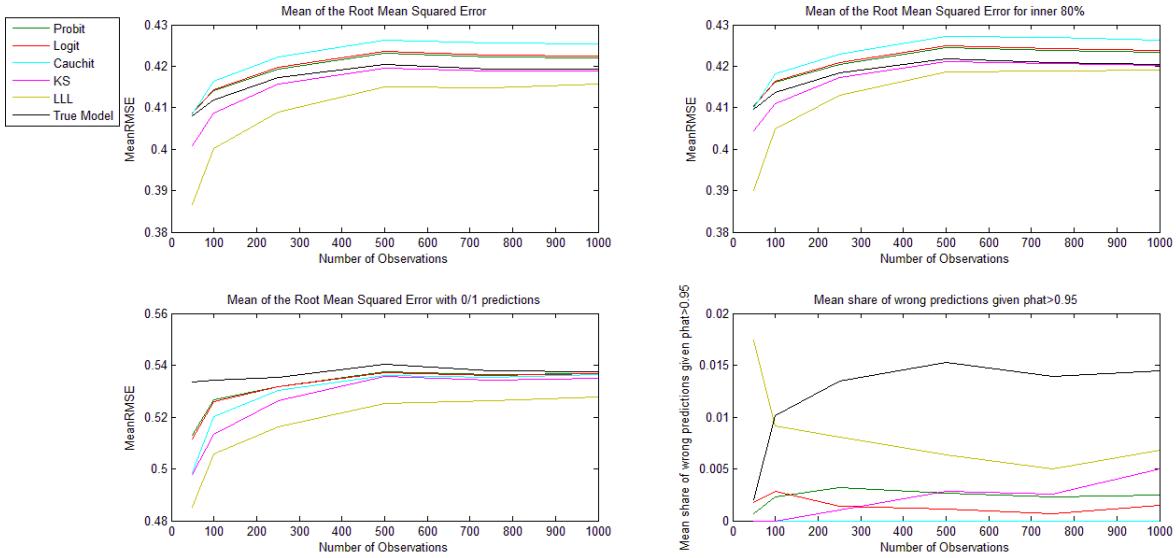
Figure 30: Distribution average ME, True Model, u is Gumbel



4.6 Setup 1 v): Bimodal errors

The following setup uses bimodal errors. With respect to the RMSE and the RMSE for the inner 80%, the semiparametric estimators perform best. Then the "true model" follows. Using the zero-one predictions the true model performs worst. The mean share of "wrong" predictions is below 2% for all estimators.

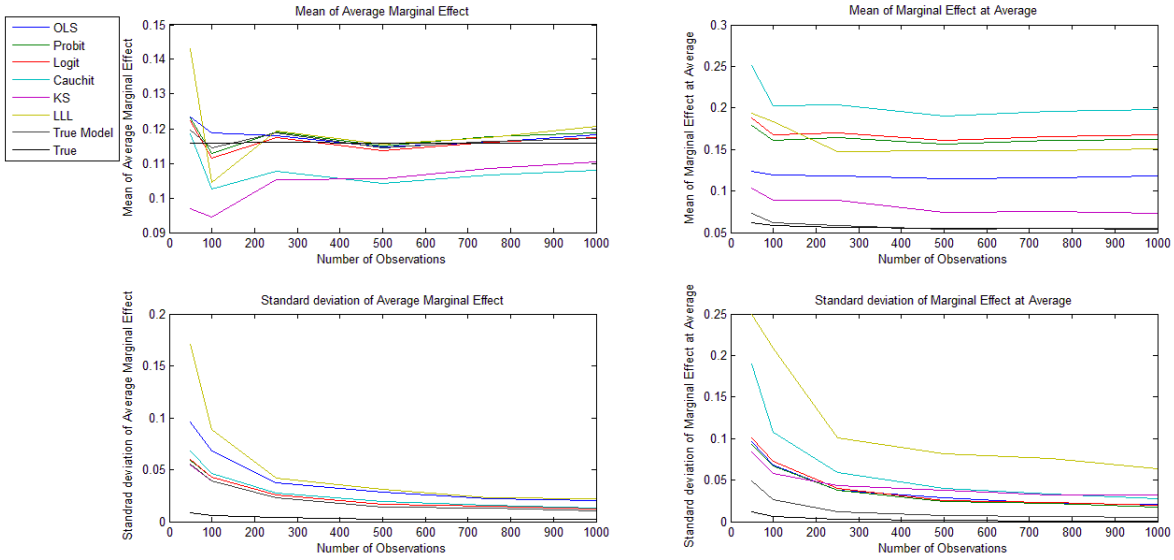
Figure 31: Performance for $\hat{y}$ given u is mixture normal



The estimators perform well in estimating the average marginal effects. Cauchit and KS perform worse than the others. As frequently observed before, KS´s estimator becomes better as the sample size increases.

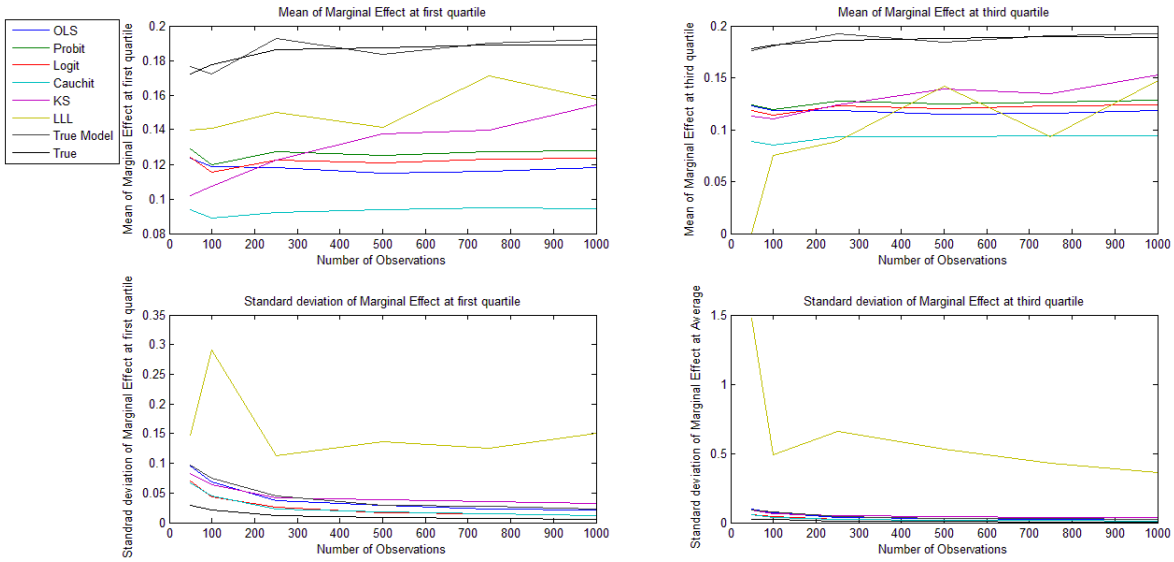
For the marginal effects at the average the only estimators with a decent performance are the ones derived from the true model and KS´s estimator.

Figure 32: Average ME and ME at average for u mixture normal



Considering the performance with respect to the estimation of the marginal effects at the quartiles, the estimator derived from the true model performs best. The performance of KS´s and the LLL estimator appears acceptable. The LLL estimator has again the property that its standard deviation does not uniformly decrease as the sample size increases.

Figure 33: ME at first quartlile and third quartile for u mixture normal



The distribution of the estimated average marginal effects again appear normal for medium and large samples.

Figure 34: Distribution average ME, Probit and Logit, u is mixture normal

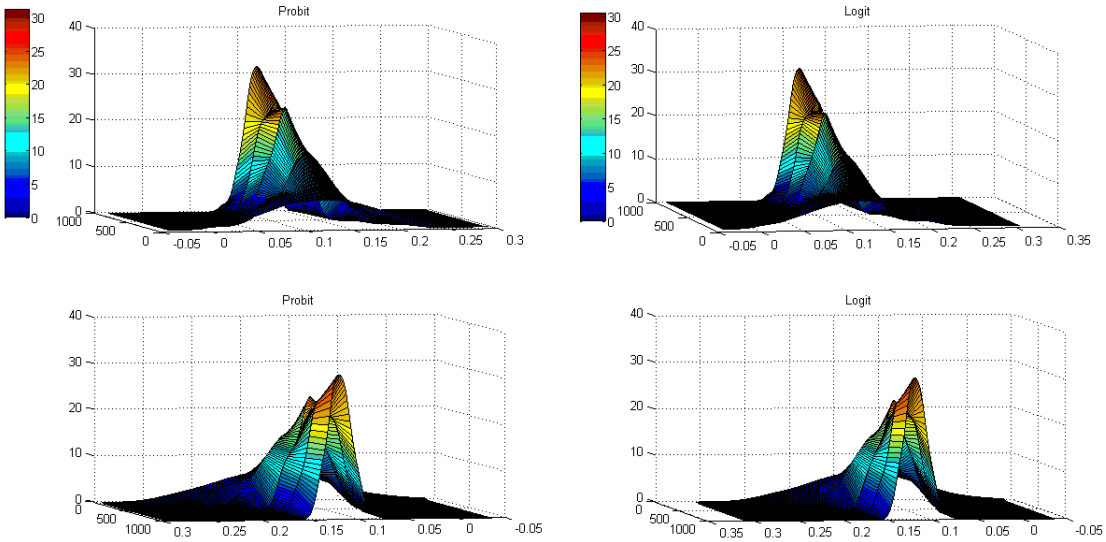


Figure 35: Distribution average ME, OLS and Cauchit, u is mixture normal

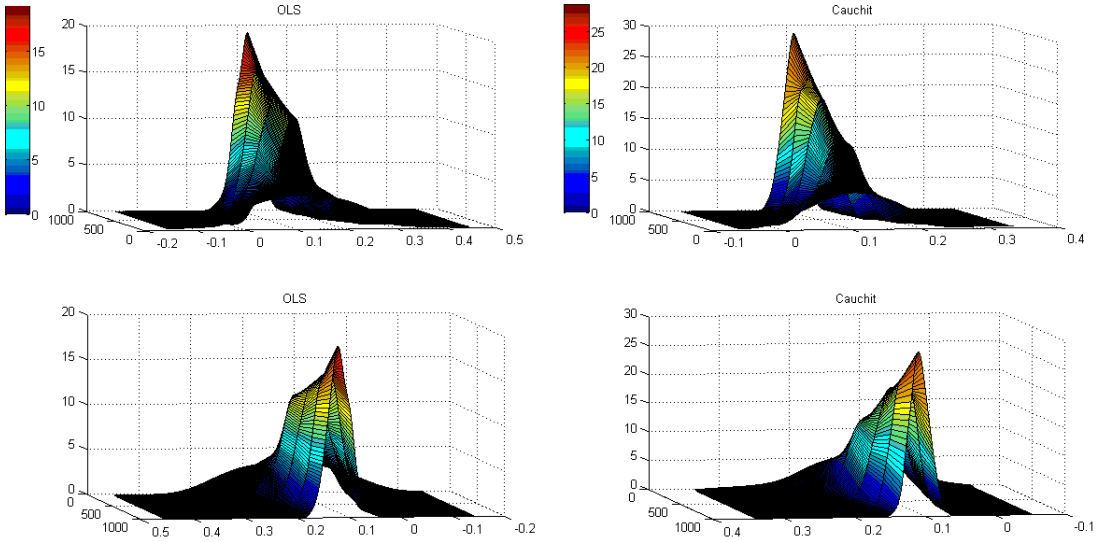


Figure 36: Distribution average ME, KS and LLL, u is mixture normal

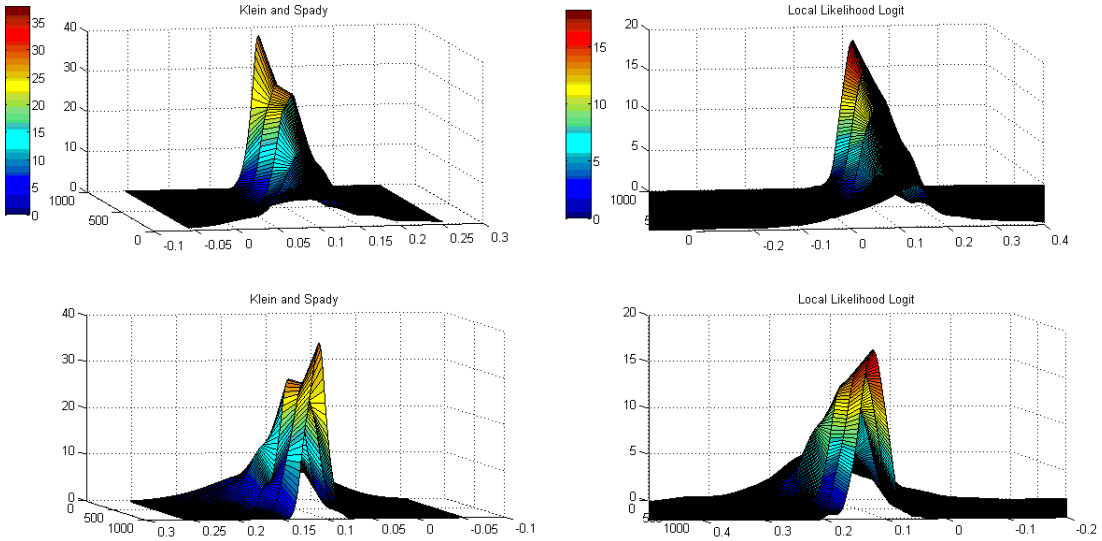
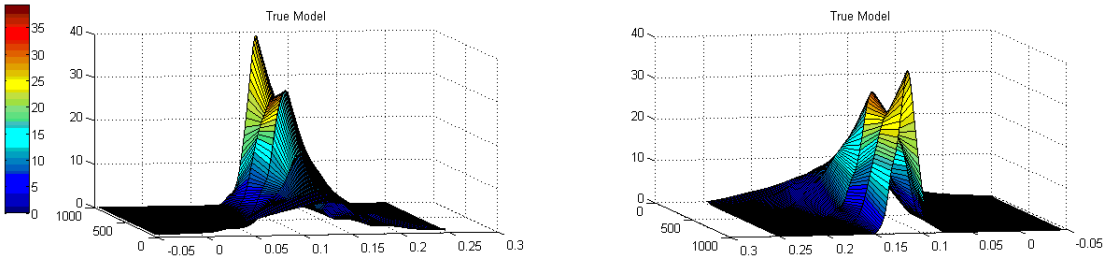


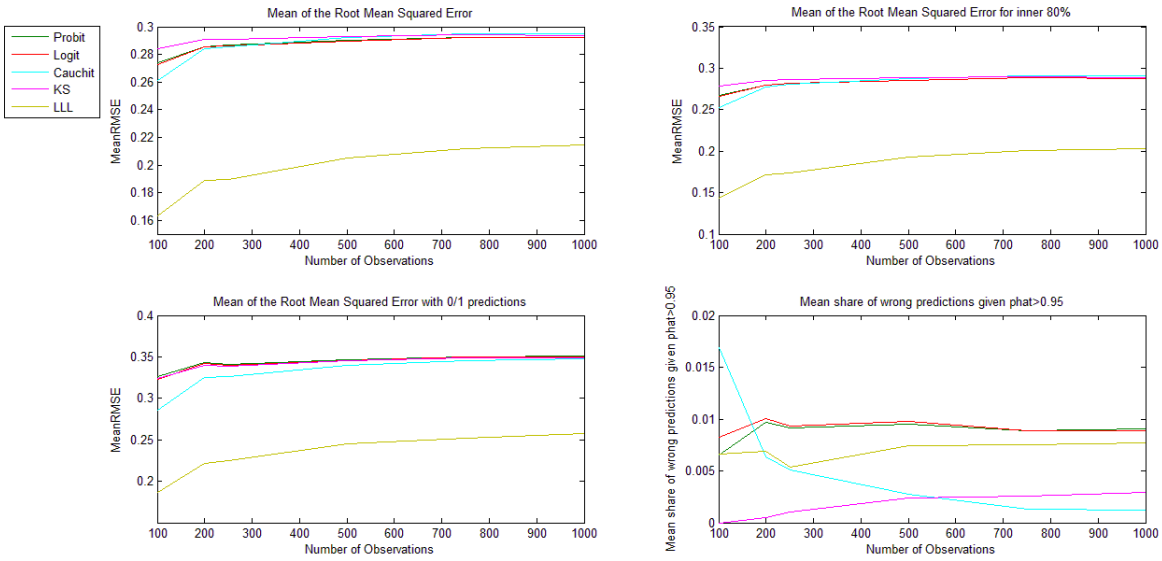
Figure 37: Distribution average ME, True model, u is mixture normal



4.7 Setup 1 vi): Many regressors

This setup is concerned with the performance of the estimators when the number of nonconstant regressors is increased from three to six. The performance of the estimators with respect to $\hat{y}$ is similar to their performance with three regressors. The LLL estimator has an even lower RMSE, whereas the remaining estimators have similar RMSE´s. The mean share of "wrong" predictions is below 2% for all estimators.

Figure 38: Performance for $\hat{y}$ given u is normal, many regressors



The performance of the estimators with respect to the average marginal effect is very similar to their performance in the three regressor case. All estimators perform acceptably. There is no substantial difference in the performance of Klein and Spady´s estimator, when varying the number of regressors. This is probably due to the semiparametric character of the estimator, which reduces the dimensionality of the nonparametric estimation. The parametric estimators, except cauchit, outperform the semiparametric ones.

Acceptable estimators for the marginal effect at the average are probit, logit, KS and LLL.

Figure 40 shows that the marginal effects at the quartiles (which are close to zero by construction), are reasonably estimated by all estimators except OLS and the LLL estimator. For the LLL estimator again twice Silverman´s plug-in estimate was used as the bandwidth choice. This might be the main reason for the relatively poor performance at the quartiles and the substantially higher standard deviation at the average marginal effect and the marginal effect at the average. An initial guess would be that the bandwidth choice should substantially increase as the number of regressors increases. Further it might be the case that the optimal bandwidth choice by cross validation has this property. Hence one conclusion of this setup is, that a careful analysis of the local likelihood logit estimator by Frölich (2006) should take its sensitivity with respect to the bandwidth choice into account.

Figure 39: Average ME and ME at average for u normal, many regressors

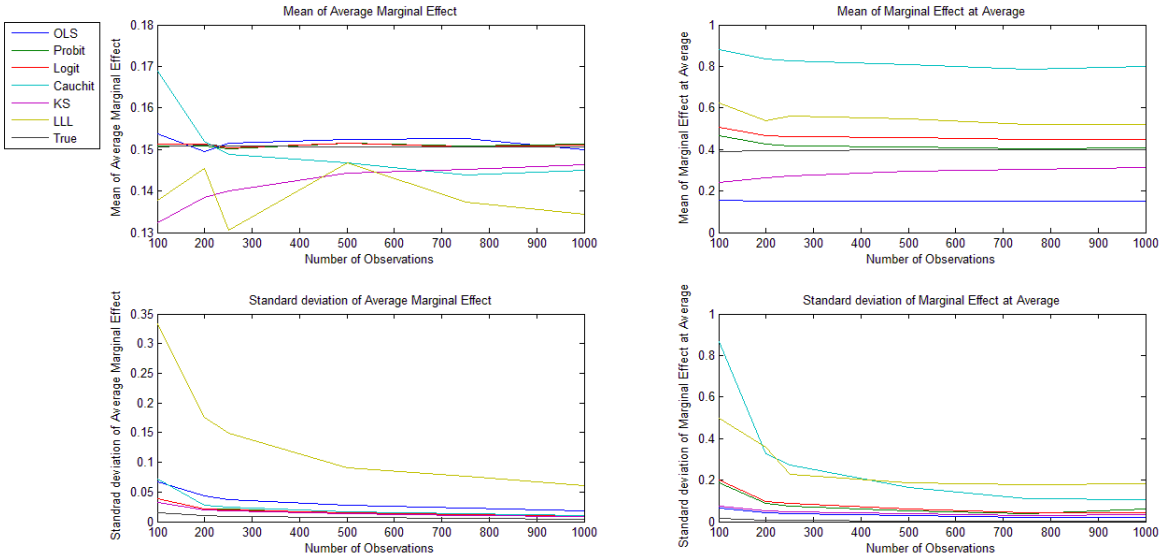
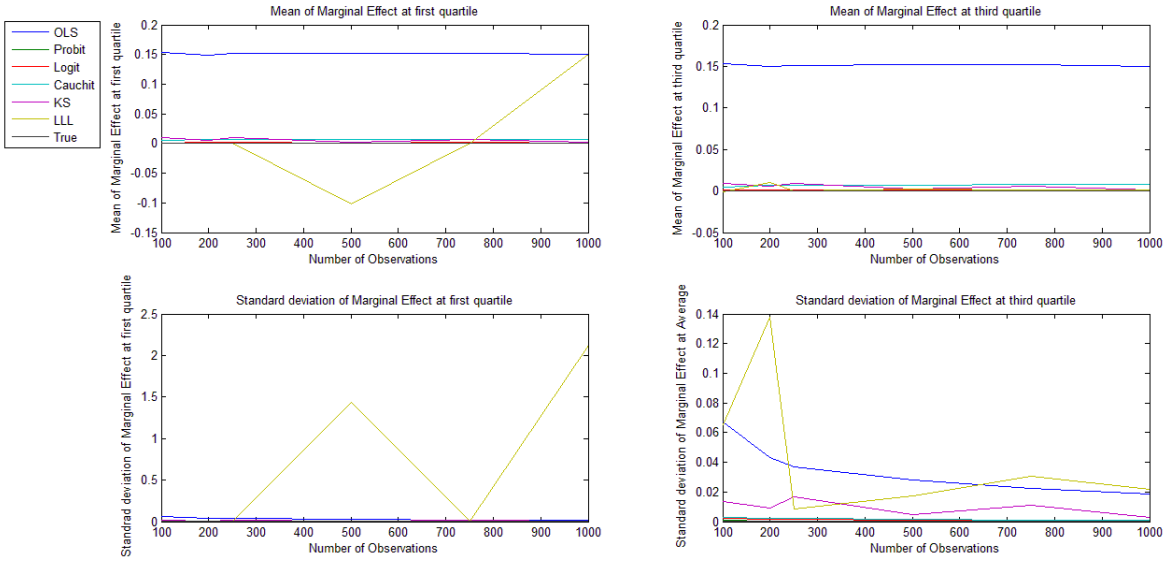


Figure 40: ME at first quartlile and third quartile for u normal, many regressors



Comparing the distributions of the estimated average marginal effects reveals the following. No sizeable changes in the distribution occur for the probit-, logit-, cauchit-, OLS- and Klein and Spadys estimator. The LLL estimator has a high variability in the estimation of the average marginal effect. This variability could be reduced by choosing higher values of the bandwidth.

Figure 41: Distribution average ME, Probit and Logit, u is normal, many regressors

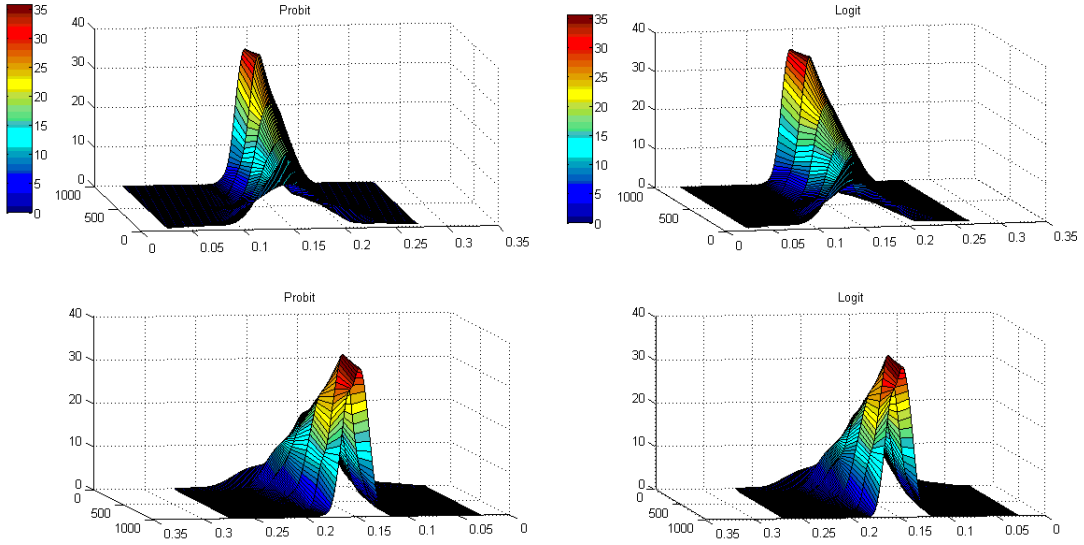


Figure 42: Distribution average ME, OLS and Cauchit, u is normal, many regressors

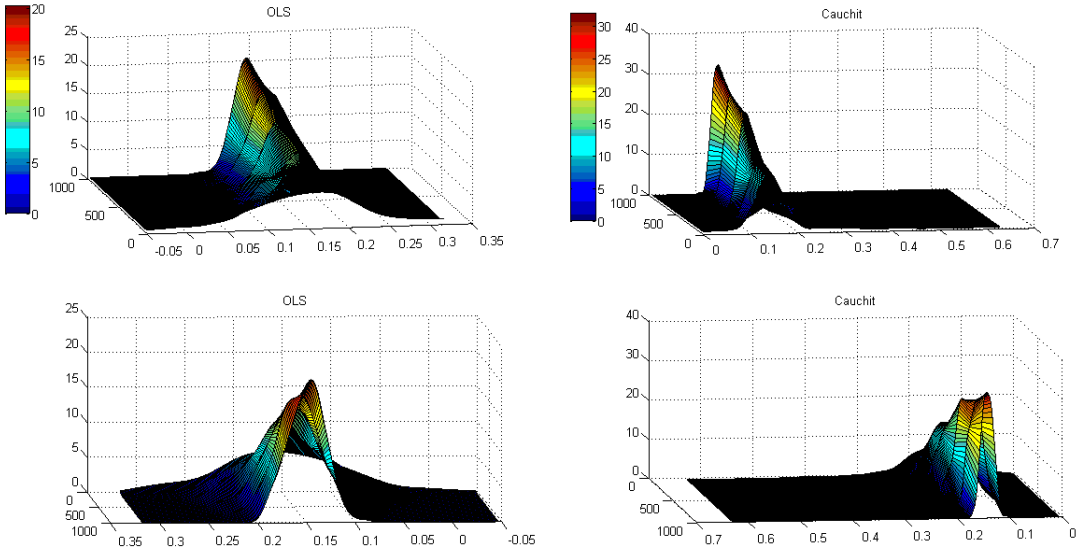
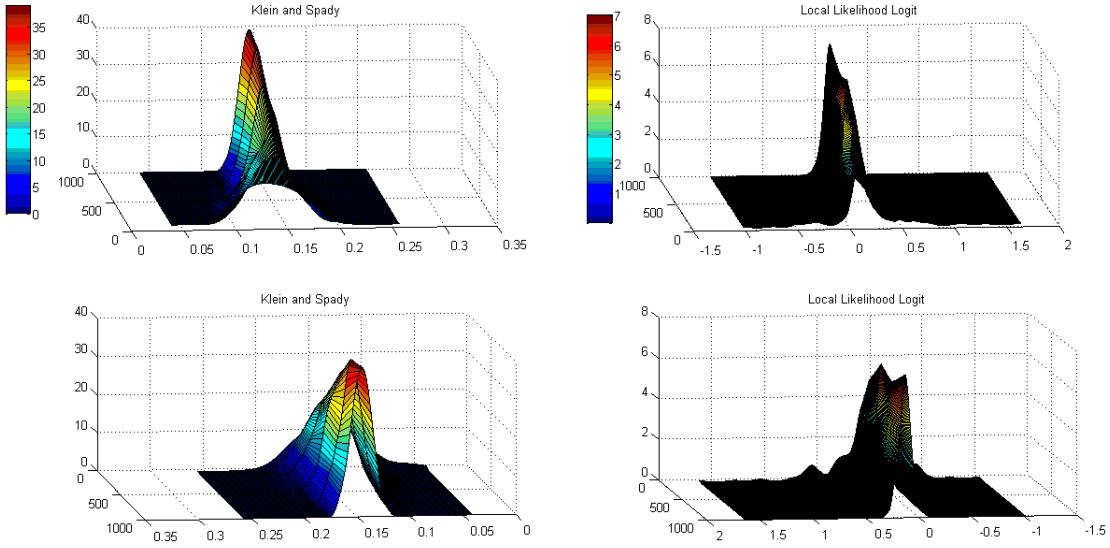


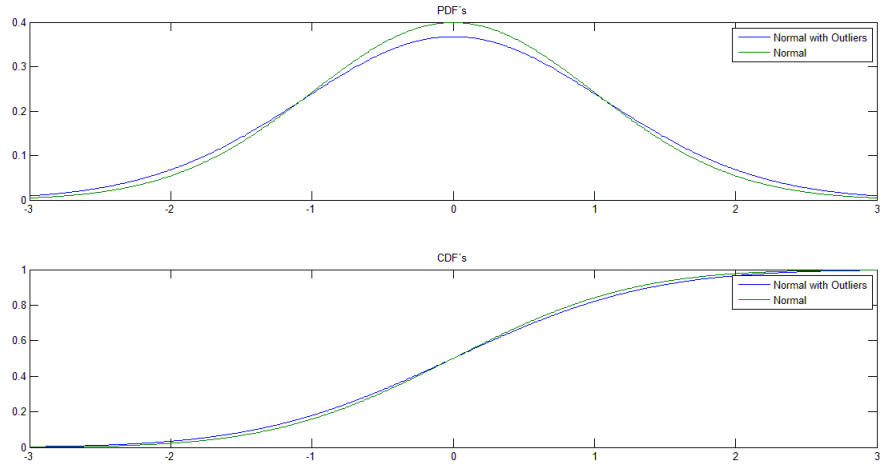
Figure 43: Distribution average ME, KS and LLL, u is normal, many regressors



4.8 Setup 1 vii): Errors with more mass in the tails

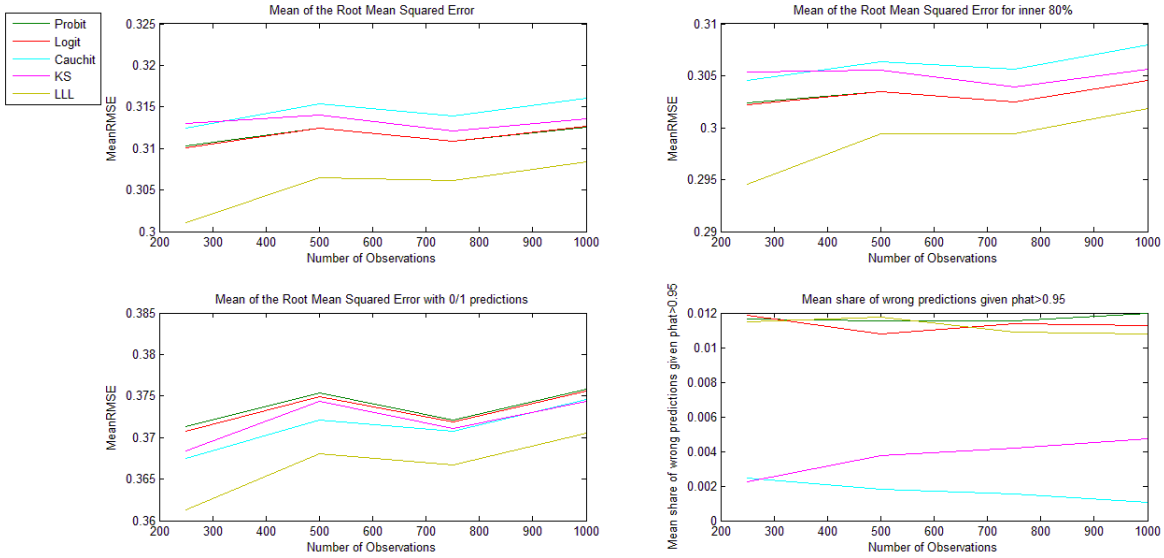
In this setup no true model was considered. Figure 44 depicts the true error distribution used in the DGP. Since there were major computational problems in the estimation of the marginal effects for the LLL estimator, this setup used sample sizes ranging from 250 to 1000.

Figure 44: True error PDF and CDF



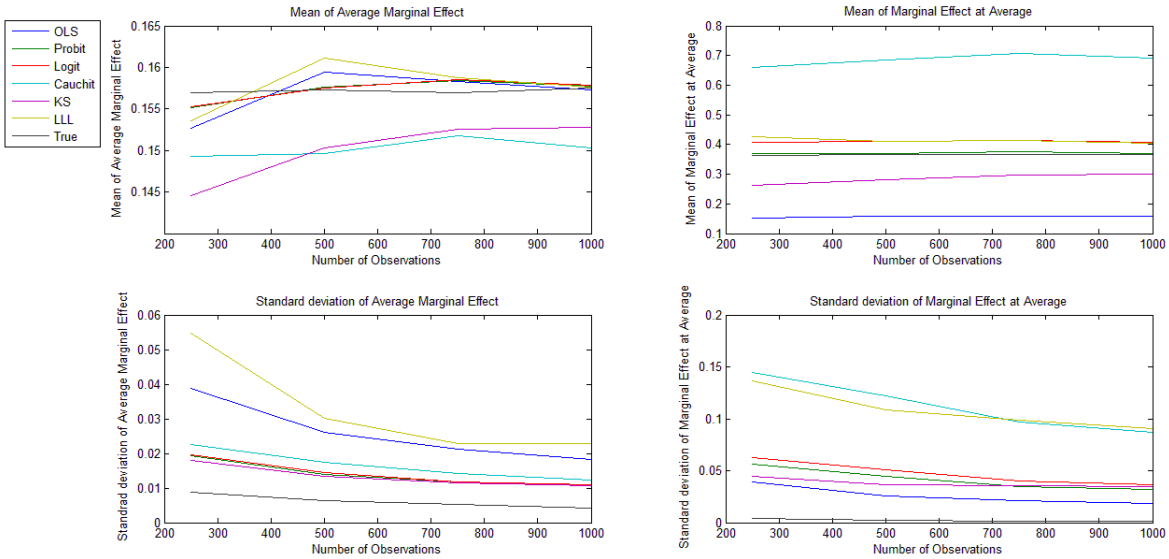
The RMSE and RMSE for the inner 80% of the sample are smallest for the LLL estimator. Klein and Spady´s- and the cauchit estimator have the highest RMSE´s. For the zero-one predictions LLL performs best, whereas logit and probit perform worst. As usual OLS is not displayed and has a substantially higher RMSE. The share of wrong predictions is at most 1.2%.

Figure 45: Performance for $\hat{y}$ given u has more mass in the tails



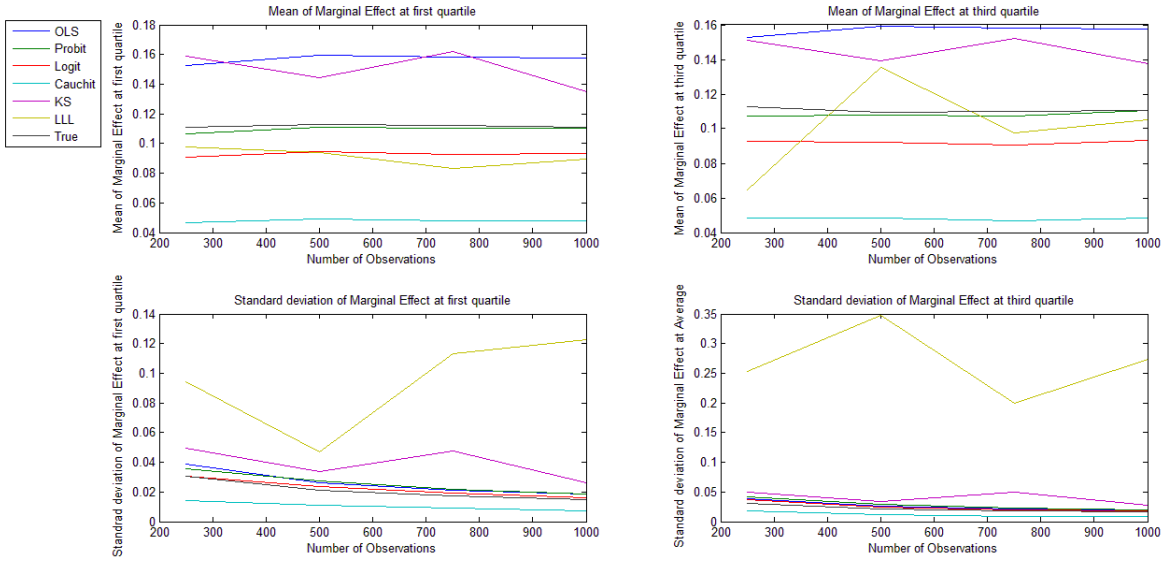
All estimators perform well in estimating the average marginal effects. KS and cauchit rank least. The marginal effect at the average is estimated well by probit. The performance of logit, LLL and KS is acceptable.

Figure 46: Average ME and ME at average for u has more mass in the tails



It seems that the probit model is able to estimate the marginal effects at the quartiles fairly well, while OLS and cauchit perfom poorly. The performance of the remaining estimators´ is acceptable. As usual, the problem that the LLL´s standard deviation not decreasing with increasing sample size remains.

Figure 47: ME at first quartlile and third quartile given u has more mass in the tails



All estimators, except the LLL estimator appear more or less normally distributed (Figure 48-50).

Figure 48: Distribution average ME, Probit and Logit given u has more mass in the tails

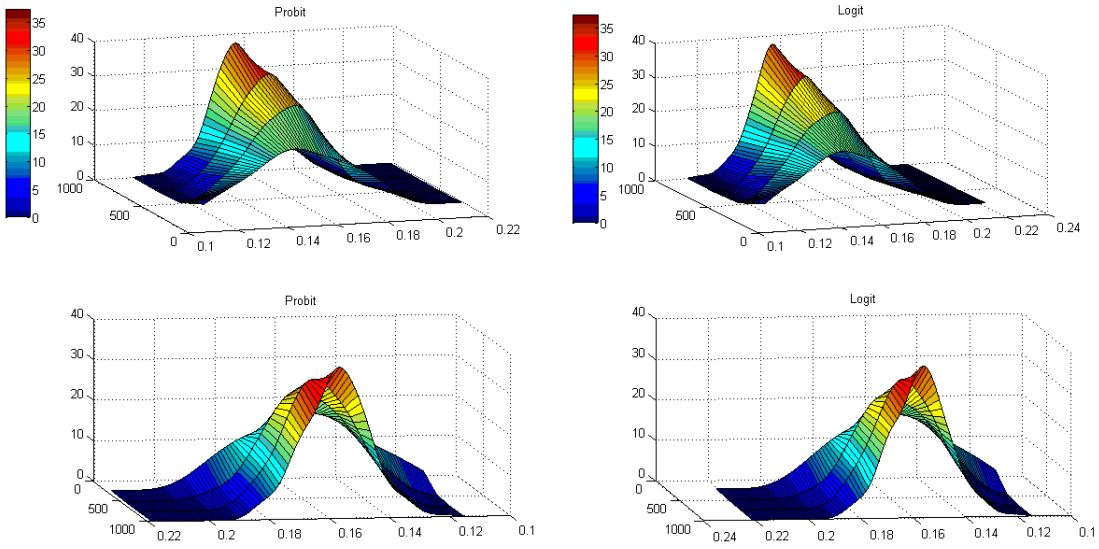


Figure 49: Distribution average ME, OLS and Cauchit given u has more mass in the tails

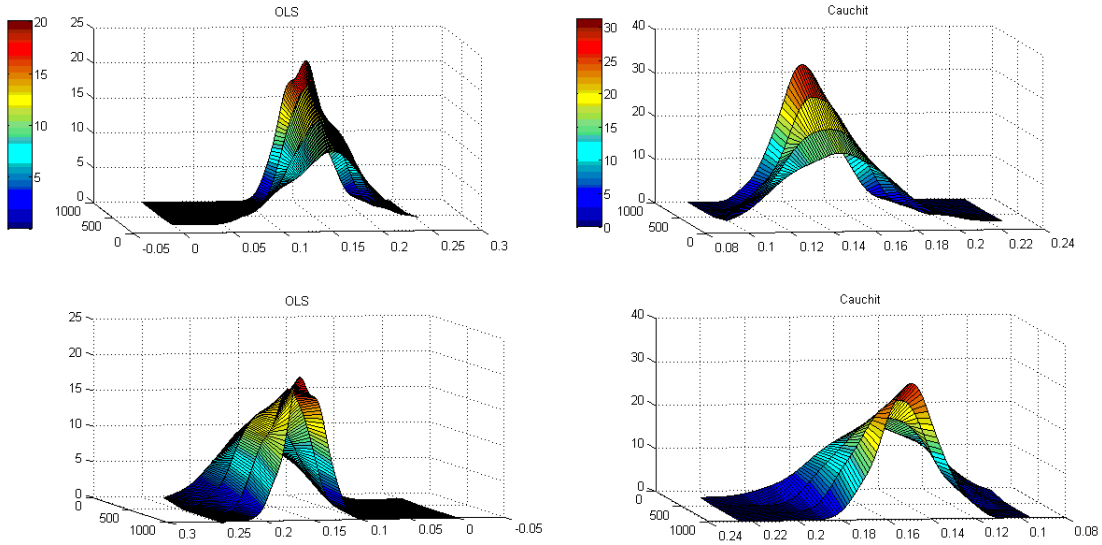
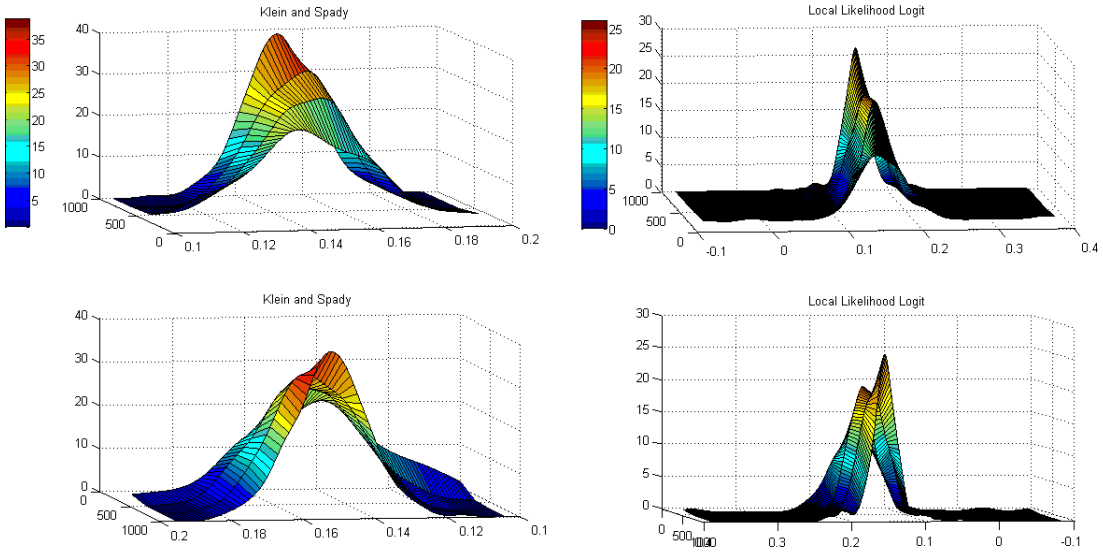


Figure 50: Distribution average ME, KS and LLL given u has more mass in the tails



Summarizing the main results of the first seven Monte Carlo setups six aspects can be perceived

1) Regarding the RMSE the local likelihood logit estimator performs best, resulting from the flexibility of the estimator. In general, Klein and Spady´s estimator ranks within the top estimators regarding the RMSE. Furthermore, in the class of the parametric estimators the true model usually performs best.

2) The share of wrong predictions is below 3% for almost all setups (except for the setup with Cauchy distributed errors).

3) With respect to the average marginal effect, the performance of all estimators is good. This supports inter alia the hypothesis that the linear probability model, which uses the OLS estimator, is useful in estimating the average marginal effect. The cauchit estimator generally performs worst (except in the case of Cauchy distributed errors). A good performance of Klein and Spady´s estimator requires a large sample.

4) The performance concerning the marginal effects at specific points varies. Usually the only model which estimates the marginal effects close to the true value is the model derived from the true distribution. Klein and Spady´s estimator comes closer to the true value as sample size grows and is often ranked second. The LLL estimators standard deviation is frequently not uniformly decreasing. The performance of the remaining parametric models is generally poor.[37]

5) The graphical inspection of the distribution of estimated average marginal effects indicates that for sufficiently large samples, the estimated average marginal effects approximately follow a normal distribution for almost all estimators.

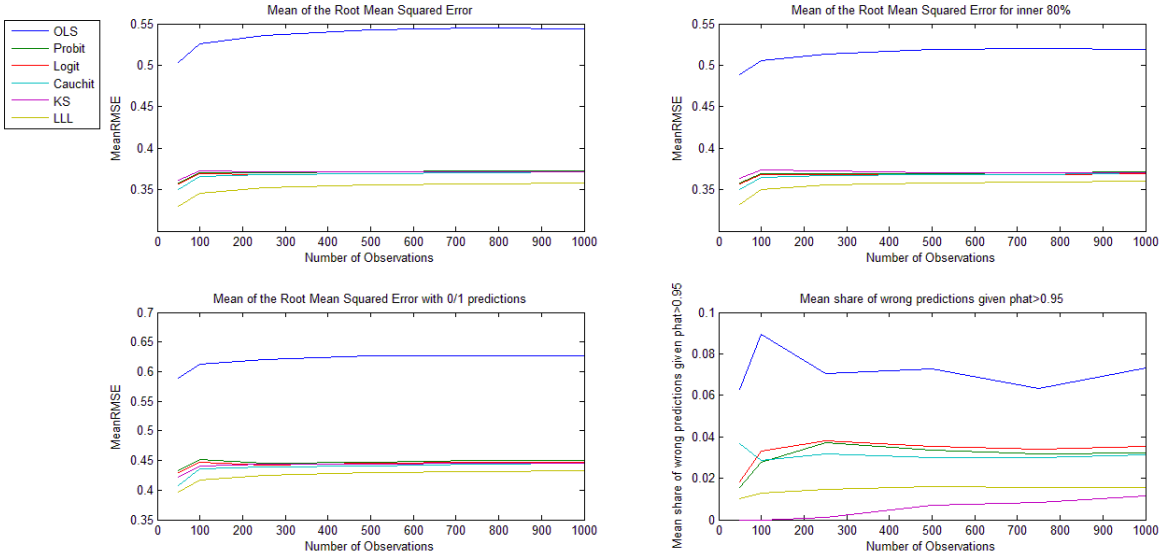
6) The performance of Klein and Spady´s estimator seems relatively unaffected by the inclusion of additional regressors. Moreover, the rule of using twice Silverman´s plug-in estimate for the bandwidth choice of the local likelihood estimator seems not reasonable when the number of regressors increases. A suggestion could be the use of cross validation or at least to increase the bandwidth "manually" as the number of regressors rises.

[37] It could be added that minor deviations from normality as present in setup vii) do not lead to extreme changes in the performance of the probit estimator.

4.9 Setup 2 i): Wrong index function ($ln(x)$ vs. x)

In setup 2 i) the functional form of the index is misspecified. The true index of the DGP uses $ln(x_3)$ whereas the estimators use x_3. Comparing the results from the well specified setup 1i) with those of the current setup reveals that every estimators RMSE is higher in the current setup, with the exception of OLS. Further the mean share of "wrong" predictions for the OLS estimator is substantially higher than usual.

Figure 51: Performance for $\hat{y}$ given wrong index function



The main conclusion from Figures 52-53 is that all estimators estimate the sign of the marginal effects correctly. Further, the performance with respect to the different marginal effects varies. The decent performance of the estimators regarding the average marginal effect does not carry over when the index is misspecified. The good performance of the probit estimator with respect to the marginal effect at the average could be due to the following. First, notice that the expected value of the third regressor is one. The true marginal effect is given by: $\frac{\partial G(x_1\beta_1+x_2\beta_2+ln(x_3))}{\partial x_3} = g(x_1\beta_1 + x_2\beta_2 + ln(x_3))\frac{1}{x_3}$ and the misspecified marginal effect equals $\frac{\partial G(x_1\beta_1+x_2\beta_2+x_3\beta_3)}{\partial x_3} = g(x_1\beta_1 + x_2\beta_2 + x_3\beta_3)\beta_3$. Due to the fact that $f(x) = ln(x)$ behaves similar to $h(x) = x$ around $x = 1$ (which can be seen by a taylor expansion) it is likely that the performance of the probit model with respect to the marginal effect at the average is due to $E(x_3) = 1$.

Moreover, no estimator is able to capture the fact that the true marginal effects at the first and the third quartile differ. As stated before, the main conclusions are: with a monotonically misspecified index the estimators estimate the sign of the marginal effects correctly, however the point estimates are not reliable.

Figure 52: Average ME and ME at average given wrong index function

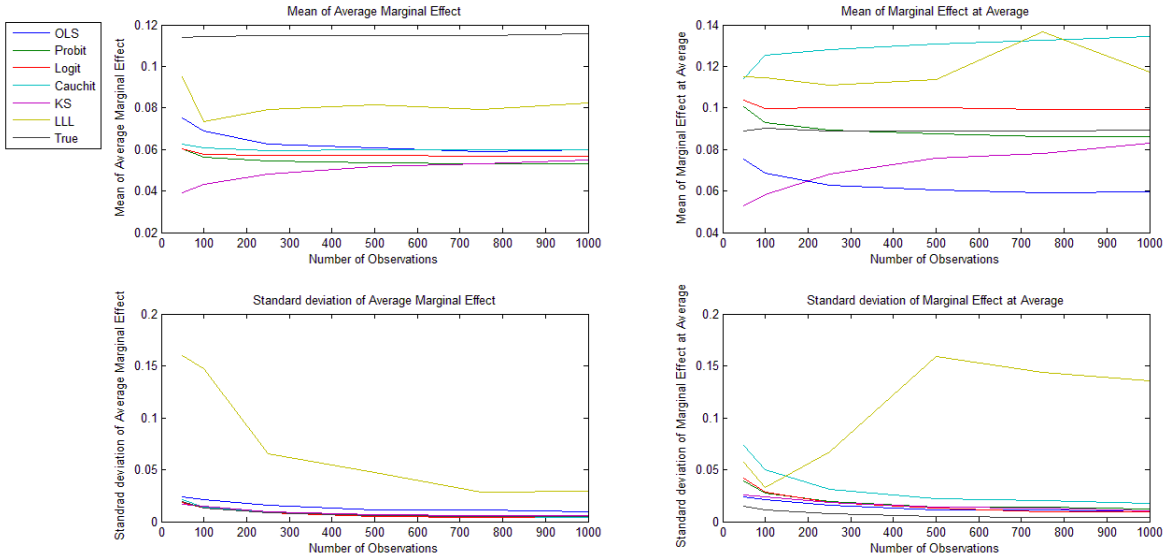
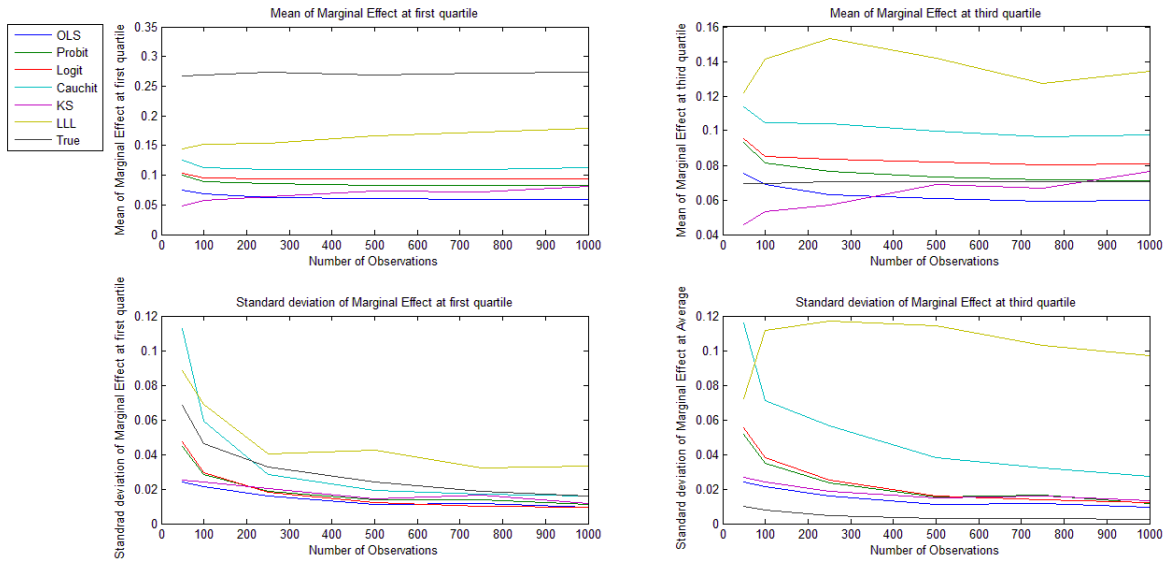


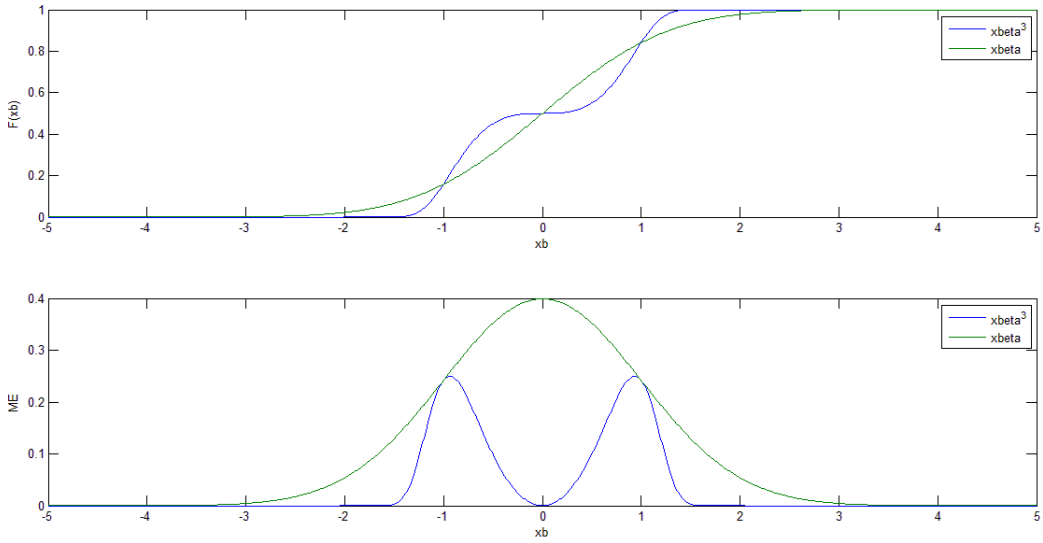
Figure 53: ME at first quartlile and third quartile given wrong index function



4.10 Setup 2 ii): Wrong index function ($(X'\beta)^3$ vs. $X'\beta$)

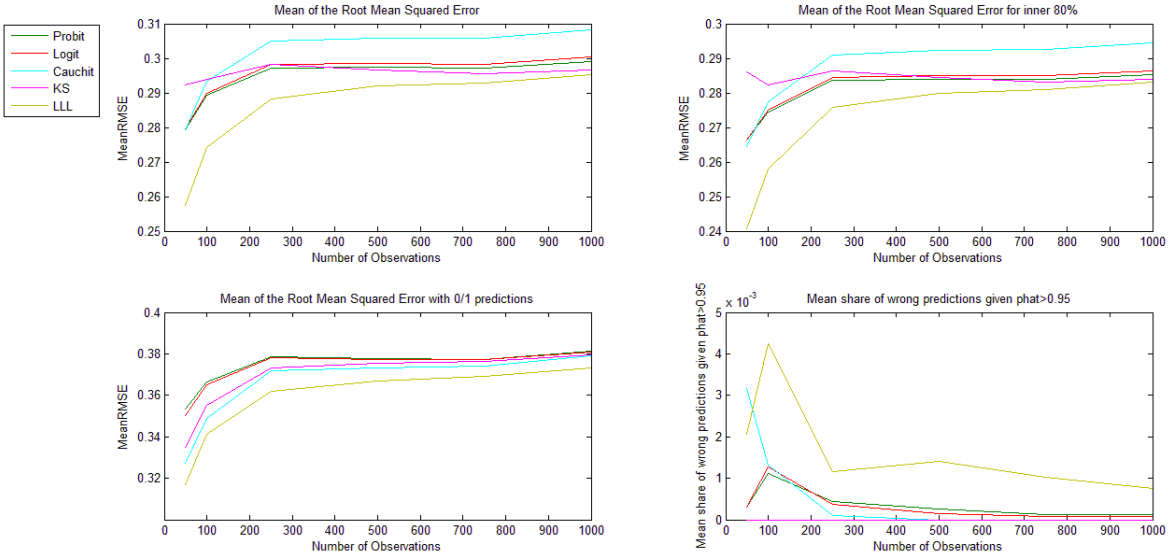
The following setup is again concerned with the importance of specifying the index function correctly. Here the correct index is $(X'\beta)^3$ but the researcher wrongly assumes that the index is given by $X'\beta$. The upper graph of Figure 54 describes the true conditional expectation $E(Y|X)$ and the one assumed, given the correct link function. The lower graph describes the true and the assumed marginal effects.

Figure 54: True versus assumed conditional expectation and marginal effects



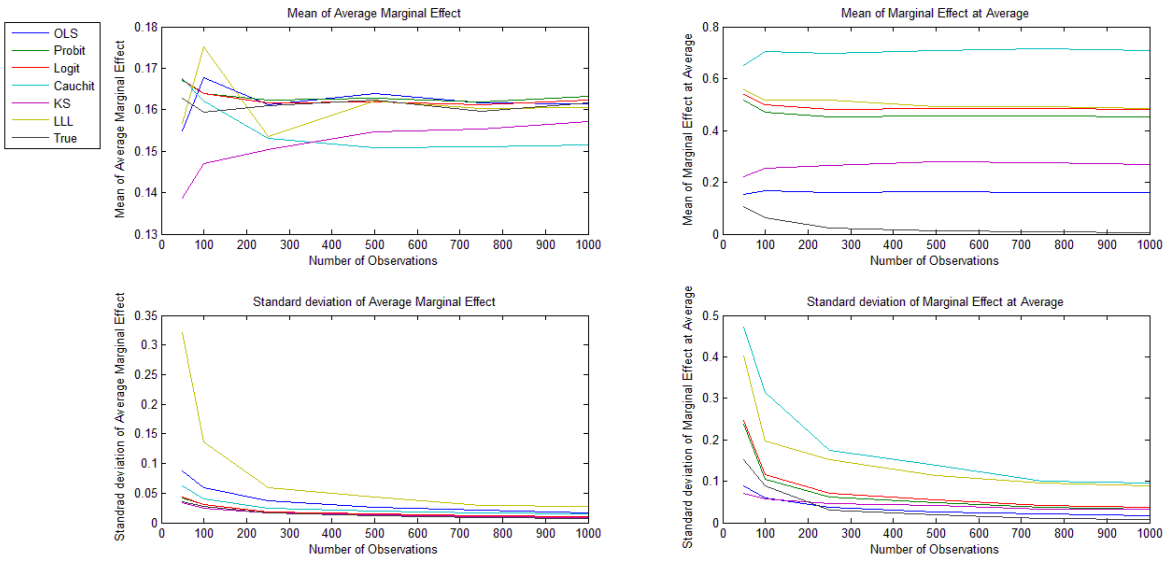
As visualized in Figure 55, the perfomance for the prediction of $\hat{y}$, does not differ substantially from the setups where a correct index function was used. The RMSE of OLS is near 0.6 and its mean share of wrong predictions is zero.

Figure 55: Performance for $\hat{y}$ given wrong index function



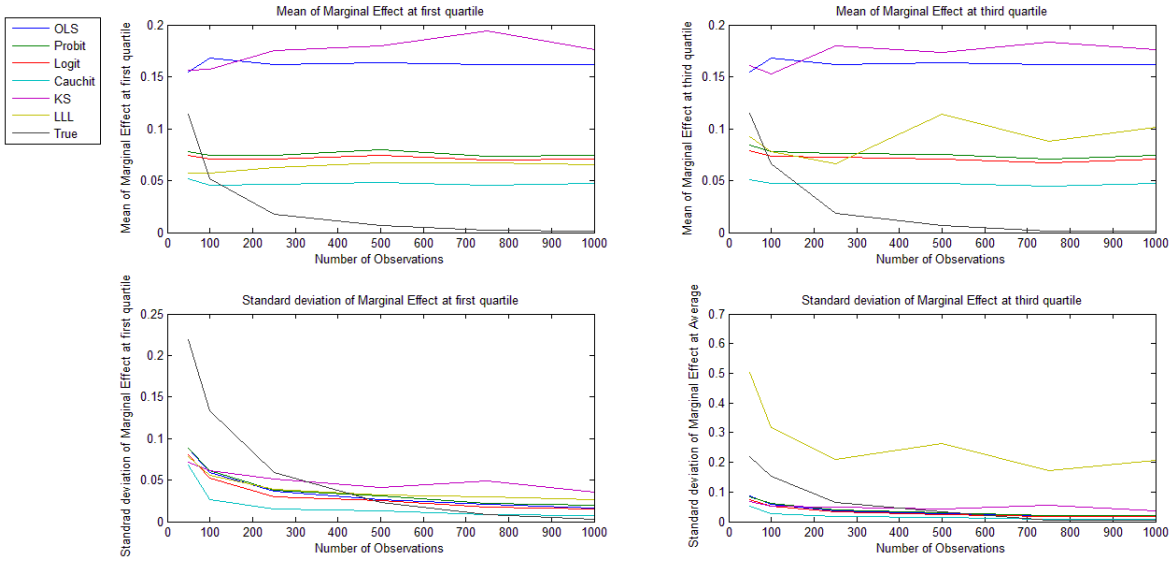
The results for the average marginal effects (Figure 56) are suprising. All estimators perform well. This can be explained by the true average marginal effect being very similar to the one in setup 1 i). Since the estimators use the same information as in setup 1 i) and a positive monotone transformation of the index, does not change the individual y_i´s the resulting estimates are identical. The performance of the estimators for the marginal effect at the average is different. All the estimators substantially overestimate the marginal effect at the average. As one can see from Figure 54 the theoretical marginal effect is close to zero. As the sample size increases the realizations of the mean of $X'\beta$ come closer to the theoretical moment and hence the true marginal effect approaches zero.

Figure 56: Average ME and ME at average given wrong index function



As for the marginal effect at the average, the performance for the marginal effects at the quartiles is poor (Figure 57). The unusual behaviour of the true marginal effects changing with sample size can be explained as follows. For large samples the realization of the quartiles of the individual regressors will be close to their theoretical values. For the setup under consideration the theoretical quartile of the index given standard normal distributed regressors becomes $Q_{0.25}(X)\beta \approx [-0.68, -0.68, -0.68] \cdot [2, -0.5, 1]' = -1.7$. The corresponding true marginal effect is zero (see Figure 54). However for small samples the realizations of the quartiles can deviate substantially and hence the true marginal effect is substantially above zero.

Figure 57: ME at first quartlile and third quartile given wrong index function



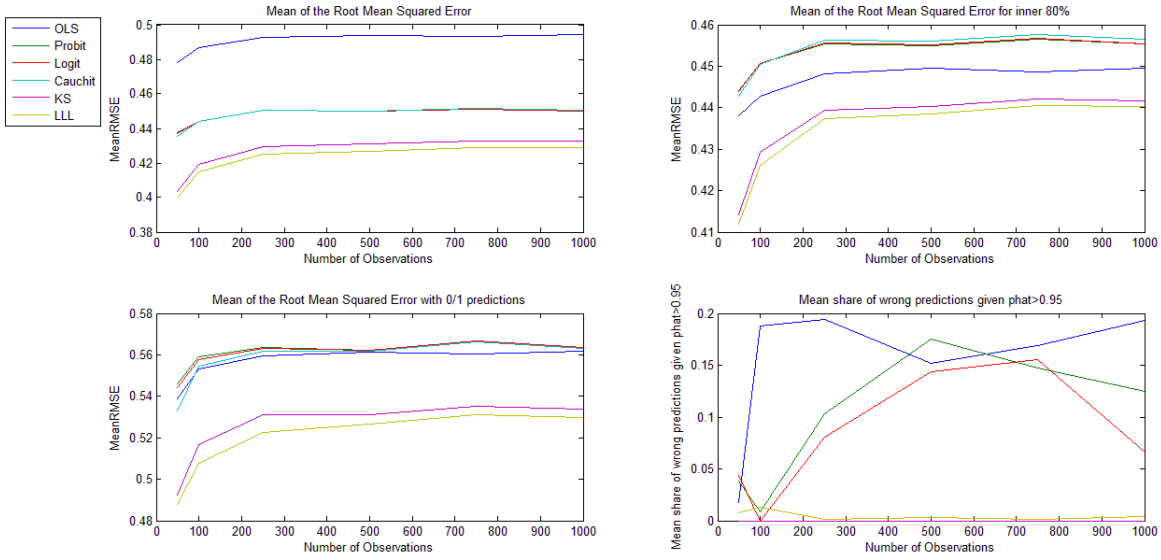
The conclusion for this setup is the following. With a misspecified index function none of the estimators under consideration is able to estimate the marginal effects at specific points consistently. However it should be noticed that the signs of the marginal effects are estimated consistently for all estimators.

4.11 Setup 2 iii): Omitted variable bias

Omitted variable bias (OVB) is a classical problem in econometrics. This setup is introduced to show the limitations of the preceding analysis. The thesis is not concerned with the various effects of omitted variable bias in nonlinear models.[38] It seems to be consensus among econometricians that nonlinearity complicates the analysis of marginal effects under omitted variable bias substantially.

The RMSE is on average higher than in the well specified setups (Figure 58). The share of "wrong" predictions yields an interesting finding. While OLS had no "wrong" predictions in most setups, the share of wrong predictions given OVB is $\approx 20\%$.[39] This result might be useful to develop a test for misspecification. As of now, this is just a preliminary idea which requires further testing. It should be possible to check the share of wrong predictions for the OLS estimator for different values of $\hat{p}$. If one finds serious deviations from the theoretical behaviour, this might indicate a serious misspecification as one resulting from omitted variable bias. This hypothesis could be tested conducting a further Monte Carlo analysis. Further it might be possible to construct an informal decision rule or even a test based on the share of wrong predictions of the OLS estimator. As stated before, this is merely an idea and further research will reveal the usefulness of the consideration.

Figure 58: Performance for $\hat{y}$ given omitted variable bias



The data generating process is such that the estimate of the coefficient in the index model is downward biased in expectation.[40]

[38] For a classical treatment see Yatchew and Griliches (1985).

[39] The only setup where OLS has a share of "wrong" predictions substantially above 5% is setup 2 i) with misspecified index.

[40] This is due to the fact that the correlation between the variable of interest and the omitted variable is 0.5 and the effect of the omitted variable on the dependent variable is negative. Further the expected correlation between the remaining regressors and the variable of interest as well as with the omitted variable is zero due to draws from independent distributions.

The results in Figure 59 support the hypothesis that omitted variable bias has unusual effects in nonlinear models. The marginal effects are estimated substantially below the true value for all estimators except OLS. Many estimators deliver negative estimates of the marginal effects, whereas the true marginal effects are positive. The decent performance of the OLS esitmator with respect to the average marginal effect raises further questions, but should not be interpreted in the sense of immunity of OLS to omitted variable bias in binary choice models.

Figure 59: Average ME and ME at average given omitted variable bias

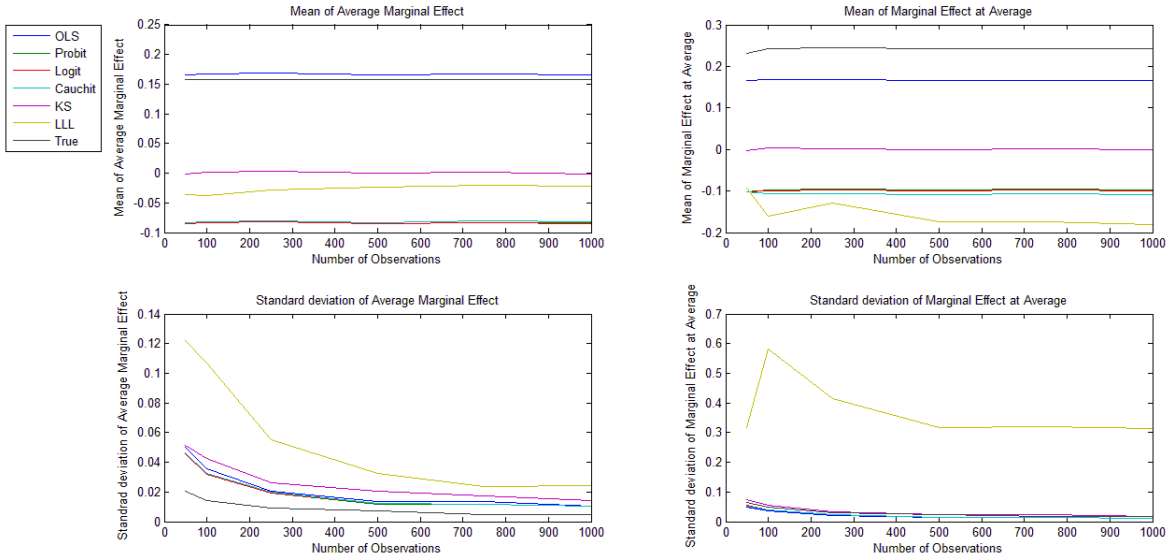
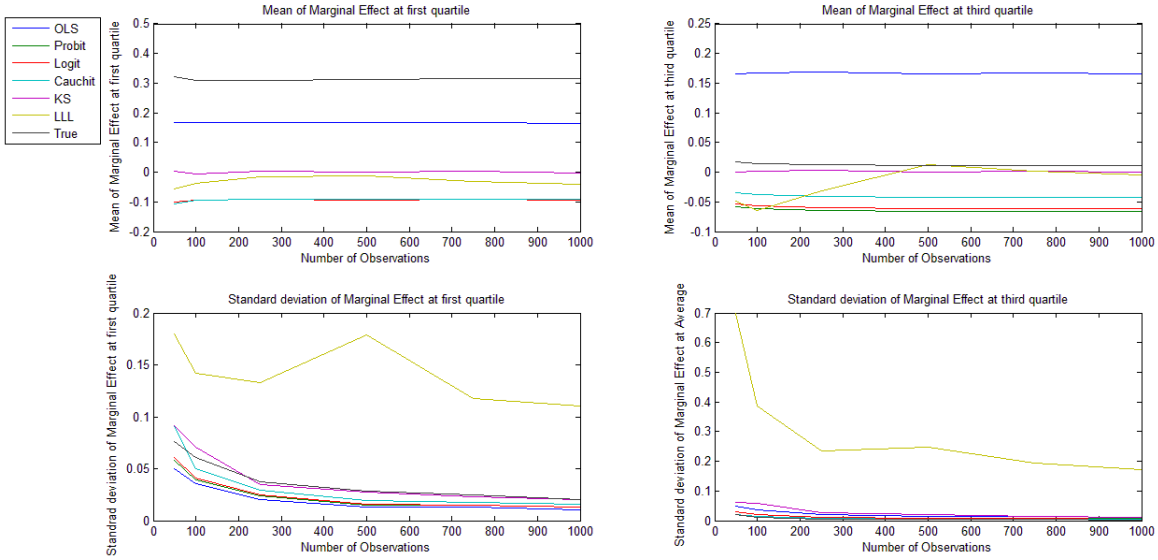


Figure 60: ME at first quartlile and third quartile given omitted variable bias



The results of the omitted variable bias setup demonstrate that the use of semiparametric estimators is not a universal remedy for all econometric problems. A further conclusion resulting from this setup is that the consequences of omitted variable bias are in general more severe than the consequences from misspecification of the functional form.

5 Possible extensions

Two major questions remain open. The first is concerned with the standard errors of the marginal effects in binary choice models. The second question focuses on the choice between parametric and semiparametric models. Two possibilities will be described to obtain estimates of the standard errors of the marginal effects, namely the "Delta Method" and "Bootstrapping". Finally, a Hausman type test is proposed to decide between parametric and semiparametric models.

5.1 Standard errors

i) Delta method

In principle, it is possible to obtain estimates of the standard errors via the so called "Delta Method".[41] Informally the Delta Method can be represented by the following statement.

> If $\sqrt{n}(\hat{\theta}_N - \theta) \xrightarrow{d} N(0, \mathbf{V})$ and the parameter space $\Theta \in \mathbb{R}^P$,
> then $\sqrt{n}(c(\hat{\theta}_N) - c(\theta)) \xrightarrow{d} N(0, (\nabla_\theta c(\theta))\, \mathbf{V}\, (\nabla_\theta c(\theta))')$.

Loosely speaking, the delta method links the asymptotic distribution of asymptotically normal distributed estimators of the parameters to the distribution of functions of these estimators. For the parametric estimators this seems a plausible way to calculate an estimate of the variance for the marginal effects. $\hat{\mathbf{V}}$ can be computed over the Hessian of the likelihood function. Further, if the researcher is interested in an individual marginal effect, he can use $c(\beta) = \frac{\partial G(X_i'\beta)}{\partial (X_i'\beta)}\beta_j$. Since β is finite dimensional and asymptotically normal with convergence rate square root n, the use of the delta method seems justified. As for the parametric estimators the asymptotic distribution for Klein and Spady´s estimator of β is available and can be estimated consistently. Owing to the fact that the function c has to be estimated and therefore is random, the theorem above is not directly applicable. In principal it should be possible to adjust the delta method, such that the asymptotic variance of the marginal effects could be worked out. A profound discussion of the delta method for nonparametric estimators is given in Ait-Sahalia (1993).

ii) Bootstrapping

A second alternative to the calculation of the standard errors is the bootstrap, which is available without relying on an estimate of the Hessian. The bootstrap method offers a simple procedure to estimate the distribution or standard errors of specific statistics. In the following

[41] For a formal statement see Wooldridge (2010, pp. 46-47).

I will give a brief summary of how bootstrapping works. This summary is mainly a restatement of some sections in Cameron and Trivedi (2005, pp. 357-365).

The starting point is an i.i.d. sample of $\{y_i, x_i\}_{i=1}^N$. The interest lies in the distribution of the estimator $\hat{\theta}$.[42] The bootstrap algorithm looks as follows.[43]

1) Resample $\{y_i^*, x_i^*\}_{i=1}^N$ from the set $\{y_i, x_i\}_{i=1}^N$ with replacement. This could result in a sample where the original $\{y_1, x_1\}$ is represented twice whereas $\{y_N, x_N\}$ is not represented at all.

2) Calculate the statistic of interest i.e. $\hat{\theta}_b$ and store it.

3) Repeat step 1) and 2) B times.

The result of this procedure is the set $\{\hat{\theta}_b\}_{b=1}^B$. From this set one can easily calculate the standard error or a kernel density estimate of $\hat{\theta}$. As a rule of thumb for B Cameron and Trivedi (2005, p. 361) suggest that $B = 384\omega$ where $\frac{\sqrt{B}(\hat{\theta}_B - \hat{\theta}_\infty)}{\hat{\theta}_\infty} \xrightarrow{d} N(0, \omega)$, whith $\hat{\theta}_\infty$ denoting the ideal bootstrap estimator with $B = \infty$. For estimation of standard errors $\omega = \frac{2+\gamma_4}{4}$ where γ_4 equals the excess kurtosis of $\hat{\theta}$. Since the main interest lies on normally distributed estimators $\gamma_4 = 0$ and therefore $B = 192$.
Following this procedure, one can easily obtain the standard errors for the estimators of the marginal effects.

5.2 A Hausman test

Until now the thesis was concerned with the performance of the different estimators given various DGP´s. Since the true DGP is unknown to the researcher, it might be useful to propose a decision rule for competing models. How could one justify the decision between competing models. Three decision criteria are prevalent in applied econometric work. First, one can generally rely on the theoretical robustness of the models under consideration, whereby the more robust model is often favoured. Second, one can rely on appropriate evidence from previous Monte Carlo studies, which compared the competing estimators under similar circumstances, such as the sample size. Those two approaches have the disadvantage that they barely take the structure of the data into account. The third decision criteria, a statistical test explicitly uses the dataset and would thus remedy this criticism. How to construct such a test, clearly depends on the models we would like to compare. From my point of view a suitable comparison would focus on the marginal effects from conventional parametric models with the ones from the semiparametric models. This comparison has the advantage of a special structure. If the parametric model is true, then the corresponding parametric etimator is efficient and Klein and Spady´s estimator is consistent but inefficient. If the parametric model is false, Klein and Spady´s estimator is still consistent, whereas the parametric estimator is not. This kind of

[42] Even though $\hat{\theta}$ can be seen as any possible estimator it might help to imagine $\hat{\theta}$ as an estimator of the marginal effects.

[43] This procedure sometimes appears under the terms empirical distribution function bootstrap, nonparametric bootstrap or paired bootstrap.

setup calls for the use of Hausman test. Frequently in Hausman tests the coefficient on the regressors of competing models are compared. However since the coefficient on the regressors do not have an intuitive explanation and different coefficients might still lead to the same conditional probabilities $p(x)$,[44] the comparison of marginal effects seems more reasonable.

The test procedure

Next, I outline the test procedure using a comparison between the logit- and Klein and Spady´s estimator. In general the comparison should be between a parametric estimator (assumed to be efficient under the null-hypothesis) and a surely consistent semiparametric estimator such as Klein and Spady´s. So if one could establish consistency and asymptotic normality for the LLL estimator, a comparison between the parametric estimators and the LLL estimator would be as well valid. The null hypothesis and the alternative are

Hypothesis:

$$\begin{aligned} &\mathrm{H}_0: \quad E(y|X) = \Lambda(X'\beta) \\ &\mathrm{H}_1: \quad E(y|X) = F(X'\beta) \quad (\text{where F(z)} \neq \Lambda(\text{z}) \text{ for some z}) \end{aligned}$$

These test statistics could be considered:

$$\begin{aligned} H_1 &= \frac{(\hat{ME}_{jLo} - \hat{ME}_{jKS})^2}{\hat{Var}(\hat{ME}_{jKS} - \hat{ME}_{jLo})} \\ H_2 &= \frac{(\hat{ME}_{jLo} - \hat{ME}_{jKS})^2}{\hat{Var}(\hat{ME}_{jKS}) - \hat{Var}(\hat{ME}_{jLo})} \\ H_3 &= \frac{(\hat{ME}_{jLo} - \hat{ME}_{jKS})^2}{\hat{Var}(\hat{ME}_{jKS})} \end{aligned}$$

It will be shown in the appendix that $H_j \overset{a}{\sim} \chi^2(1)$ under H0 for $j = 1, 2, 3$.

Critical value:

Depending on the level of the type I error (α), the critical values are 3.84 ($\alpha = 0.05$), 6.64 ($\alpha = 0.01$) or 10.83 ($\alpha = 0.001$).

Decision Rule:

Discard the null hypothesis if the value of the test statistic is above the critical value. This indicates evidence against the parametric model and suggests to choose Klein and Spady´s estimator.

Intuition:

If H_0 is true it is likely that the estimates of the marginal effects are very similar in both specifications. This is due to the fact that both estimators are consistent under H_0. If the alternative H_1 is true, the fact that one estimator is consistent and the other is not, suggests

[44] As mentioned before, this is due to the two sided dependence of $p(x)$ on the choice of the link function and the index function $X'\beta$. Differences in the β´s might be offset by differences in the link function.

that the difference between the two estimators is different from zero. The distribution of the test statistics can be made plausible by the following consideration. As will be shown in the appendix, the estimators of the marginal effects are asymptotically normally distributed. Additionally, if one normalizes the normally distributed estimators of the marginal effects by their standard deviation, the result is a standard normal distribution. Due to the fact that the square of a standard normal distributed random variable is a chi-square distributed random variable the test statistics are asymptotically chi-square distributed.[45] The three proposed test statistics merely differ in the estimation of the variance of the difference of the estimators for the marginal effects. The test statistic H_1 uses a standard estimator for the variance. H_2 utilizes that Hausman (1978, pp. 1253-1254) showed that the asymptotic variance of the differences has the following form, $Var(\hat{ME}_{Lo} - \hat{ME}_{KS}) = Var(\hat{ME}_{KS}) - Var(\hat{ME}_{Lo})$. The test statistic H_3 has only a weak theoretical justification. The idea is that the marginal effects from Klein and Spadys estimator converge at rate $\sqrt{Nh^3}$, whereas the parametric marginal effects converge at rate $\sqrt{N}$. Therefore the variance estimator $\hat{Var}(\hat{ME}_{Lo}) \xrightarrow{p} 0$, much faster than $\hat{Var}(\hat{ME}_{KS})$. Hence the effect of $Var(\hat{ME}_{Lo})$ might be close to zero.

Critique:
An appropriate analysis of the test properties was beyond the scope of the thesis and thus not conducted. Before such an analysis is not carried out, it would not be reasonable to apply the test in practice. Hence, one part of my future research will be devoted to conduct an extensive size and power analysis of the proposed test statistics and compare the test to already existing ones (see for example Pagan and Ullah (1999, pp. 141-150)).

6 Conclusions

It was stated in the introduction that the goal of the thesis is to give practical guidance for the selection of binary choice estimators in applied work. The Monte Carlo study suggests that there is no unique optimal estimator for all setups. An applied researcher should answer two questions before deciding for any of the proposed estimators. First, what is the quantity of interest? Second, is the sample large enough to use a semiparametric estimator given the number of regressors? If the researcher is interested in the average marginal effect of a change in a specific variable, the Monte Carlo study suggests that each of the estimators can be used. Even the crude linear probability model gives reasonable estimates. If the researcher is interested in the marginal effects at specific points or the distribution of marginal effects and the sample size is sufficiently large, the Monte Carlo study suggests to use the semiparametric estimators. At this point the suggestion would further be to rely on Klein and Spady´s estimator. The local likelihood logit estimator was not always found to be consistent. But it should in general be kept in mind that the lack of using cross validation for the semiparametric estimators might underestimate their performance. The second question

[45] The test statistic can be easily extended to jointly test the difference of all marginal effects, as it is done in the appendix.

relates to the appropriateness of the sample size. As especially the last setup, concerned with omitted variable bias, suggests it seems more important to include all relevant variables than to decide between parametric or semiparametric estimators. Given a sufficiently large sample the researcher could compute the marginal effects from a parametric model and informally compare the values with those from a semiparametric model. If the values are close, one could as a rule of thumb use the parametric estimator. If they depart substantially in large samples, the suggestion would be to use the semiparametric estimator. If the sample is relatively small one should use a parametric estimator to estimate the average marginal effect. Finally, researchers interested in the marginal effects at specific points should be aware of the small sample bias of the semiparametric estimators as well as the substantial inconsistency of misspecified parametric estimators.

Appendix

Derivations and Proofs

The following four theorems will be used. They are taken from Wooldridge (2010, pp. 37-47).

a) Weak law of large numbers: Let $\{\mathbf{w}_i : i = 1, 2, ...\}$ be a sequence of independent, identically distributed $G \times 1$ random vectors such that $E(|w_{ig}|) < \infty$, $g = 1, ..., G$. Then the sequence satisfies the **weak law of large numbers (WLLN):** $N^{-1} \sum_{i=1}^{N} \mathbf{w}_i \xrightarrow{p} \mu_{\mathbf{w}}$, where $\mu_{\mathbf{w}} = E(\mathbf{w}_i)$

b) Slutsky´s theorem: Let $g : \mathbb{R}^K \to \mathbb{R}^J$ be a function continuous at some point $\mathbf{c} \in \mathbb{R}^{\mathbb{K}}$. Let $\{\mathbf{x}_N : N = 1, 2, ...\}$ be a sequence of $K \times 1$ random vectors such that $\mathbf{x}_N \xrightarrow{p} \mathbf{c}$. Then $g(\mathbf{x}_N) \xrightarrow{p} g(\mathbf{c})$ as $N \to \infty$. In other words, $plim(g(\mathbf{x}_N)) = g(plim(\mathbf{x}_N))$ if $g(\cdot)$ is continuous at plim $\mathbf{x}_N$.

c) Delta method: Let $\{\hat{\theta}_N : N = 1, 2, ...\}$ be a sequence of estimators of the $P \times 1$ vector $\theta \in \Theta$. Suppose that $\sqrt{N}(\hat{\theta}_N - \theta) \xrightarrow{d} N(0, \mathbf{V})$, where $\mathbf{V}$ is a $P \times P$ positive semidefinite matrix, and let c: $\mathbf{\Theta} \to \mathbb{R}^Q$ be continuously differentiable function on the parameter space $\Theta \subset \mathbb{R}^P$, where $Q \leq P$, and assume that θ is in the interior of the parameter space, then $\sqrt{N}(c(\hat{\theta}_N) - c(\theta)) \xrightarrow{d} N(0, (\nabla_\theta c(\theta))\, \mathbf{V}\, (\nabla_\theta c(\theta))')$, where $\nabla_\theta c(\theta)$ denotes the gradient of c.

d) Asymptotic equivalence lemma: Let $\{\mathbf{x}_N\}$ and $\{\mathbf{z}_N\}$ be sequences of $K \times 1$ random vectors. If $\mathbf{z}_N \xrightarrow{d} \mathbf{z}$ and $\mathbf{x}_N - \mathbf{z}_N \xrightarrow{p} \mathbf{0}$, then $\mathbf{x}_N \xrightarrow{d} \mathbf{z}$.

i) The probability limit of the average marginal effect and the marginal effect at the average

In the following I will derive the probability limit of the average marginal effect and the marginal effect at the average treating x as a $1 \times k$ random vector and β as known.

The true marginal effect at the average is given by

$$\frac{\partial E(y_i|\bar{x})}{\partial x_{ij}} = g(\bar{x}\beta)\beta_j$$

From the weak law of large number we know that sample averages converge to their expected value. Hence

$$plim(\bar{x}\beta) \overset{WLLN}{=} E(x_i\beta)$$

Further we know by Slutsky´s theorem that

$$plim(g(\bar{x}\beta)\beta_j) \overset{Slutsky}{=} g(plim(\bar{x}\beta))\beta_j = g(E(x_i\beta))\beta_j$$

If we directly apply the WLLN to the true average marginal effect, we receive

$$\frac{1}{n}\sum \frac{\partial E(y_i|x_i)}{\partial x_{ij}} = \frac{1}{n}\sum \frac{\partial G(x_i\beta)}{\partial x_{ij}} = \frac{1}{n}\sum g(x_i\beta)\beta_j$$

$$plim(\frac{1}{n}\sum g(x_i\beta)\beta_j) \overset{WLLN}{=} E(g(x_i\beta))\beta_j$$

Due to the fact, that $g(E(x_i\beta)\beta_j \neq E(g(x_i\beta))\beta_j$ the two objects of interest have a different probability limit. To get an idea concerning the difference, knowledge of the shape of g is needed. If g is either concave or convex Jensens inequality can be employed. For the quasiconcave normal distribution it seems to be that "the concave part dominates" and $g(E(x_i\beta)\beta_j \geq E(g(x_i\beta))\beta_j$ which means that the marginal effect at the average is larger than the average marginal effect.

ii) Inverse transform sampling

In the cases were no pseudo random number generator was implemented in matlab, I generated the random numbers with the inverse transform sampling method. Suppose we want to draw pseudo random numbers from a random variable X with CDF F_X. Further we have already produced draws from a uniform distribution from zero to one (denoted by u_i) using matlabs built in function. The inverse transform sampling method then works as follows.

a) Construct the inverse of F_x, denoted by F_x^{-1}.
b) Evaluate the inverse of F_x at u_i.
c) The pseudo random draw $X_i = F_x^{-1}(u_i)$.

For details see Rinne (2003, p. 209).

iii) The asymptotic normal distribution of the marginal effects

Establishing the asymptotic normality of the marginal effects in parametric models, given that the coefficient estimates are normally distributed is straightforward. We know that $\sqrt{N}(\hat{\theta}_N - \theta) \xrightarrow{p} N(0, \mathbf{V})$, by direct application of the delta method we receive $\sqrt{N}(c(\hat{\theta}_N) - c(\theta)) \xrightarrow{d} N(0, (\nabla_\theta c(\theta))\,\mathbf{V}\,(\nabla_\theta c(\theta))')$. Now setting $c(\theta) = g(X_i'\theta)\theta = ME_{x_i}$ shows that the estimator of the marginal effects are asymptotically normally distributed, if the coefficient estimators are normally distributed. For Klein and Spady´s estimator the discussion is more elaborate. Pagan and Ullah (1999, p. 177) describe the conditions for the partial derivative

estimator being asymptotically normal. Moreover, Pagan and Ullah (1999, p. 165) show that the difference between the finite difference- and the partial derivative estimator is $O(h)$. Formally, we know that under some conditions $\sqrt{Nh^3}(\hat{ME}_{PD} - ME) \xrightarrow{d} N(0, \mathbf{V})$. Further if we require $h \to 0$ as $N \to \infty$, one can conclude that $\hat{ME}_{FD} - \hat{ME}_{PD} = o(h^{1+\epsilon})$ from the fact that $\hat{ME}_{FD} - \hat{ME}_{PD} = O(h)$. Since $h \to 0$ we have $\hat{ME}_{FD} - \hat{ME}_{PD} = o(1)$, which is equivalent to $\hat{ME}_{FD} - \hat{ME}_{PD} \xrightarrow{p} \mathbf{0}$. Collecting both results, we can apply the asymptotic equivalence lemma. If $\sqrt{Nh^3}(\hat{ME}_{PD} - ME) \xrightarrow{d} N(\mathbf{0}, \mathbf{V})$ and $\sqrt{Nh^3}\hat{ME}_{FD} - \sqrt{Nh^3}\hat{ME}_{PD} \xrightarrow{p} \mathbf{0}$, then $\sqrt{Nh^3}(\hat{ME}_{FD} - ME) \xrightarrow{d} N(\mathbf{0}, \mathbf{V})$.

iv) The derivation of the asymptotic distribution of the Hausman test statistic

The following formulas are given in terms of the marginal effects of all variables. Hence using the test statistics as a joint test is possible. The test statistics proposed in section 5.2. can be deduced by replacing the marginal effect vector and the covariance matrix of the marginal effects by the marginal effects of a variable and its variance. The derivation of the asymptotic distribution of the test statistics is close to the one in Hausman (1978, p. 1256). The discussion of the asymptotic properties of the test statistic is preliminary and merely heuristic. The main idea is to show that, the difference of the estimators is normally distributed. Then it follows from a standard relation between the normal- and the chi-square distribution that $H = (\hat{ME}_{Lo} - \hat{ME}_{KS})Var(\hat{ME}_{Lo} - \hat{ME}_{KS})^{-1}(\hat{ME}_{Lo} - \hat{ME}_{KS})'$ is asymptotically distributed as χ^2_K.

Initially we know from the discussion above that

$$\sqrt{Nh^3}(\hat{ME}_{KS} - ME) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{KS}) \; and \; \sqrt{N}(\hat{ME}_{Lo} - ME) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{Lo})$$

Since both $\hat{ME}_{KS}$ and $\hat{ME}_{Lo}$ are consistent under the null-hypothesis, $plim(\hat{ME}_{Lo} - \hat{ME}_{KS}) = 0$. With a central limit theorem we can show that $\sqrt{Nh^3}(\hat{ME}_{Lo} - \hat{ME}_{KS}) \xrightarrow{d} N(0, \Sigma)$. Since it is established that $(\hat{ME}_{Lo} - \hat{ME}_{KS})$ is asymptotically normal it follows directly that $H = (\hat{ME}_{Lo} - \hat{ME}_{KS})Var(\hat{ME}_{Lo} - \hat{ME}_{KS})^{-1}(\hat{ME}_{Lo} - \hat{ME}_{KS})'$ is χ^2_K. Reducing the formula for H to a univariate marginal effect and replacing $Var(\hat{ME}_{jLo} - \hat{ME}_{jKS})$ by a consistent estimator $\hat{Var}(\hat{ME}_{jKS} - \hat{ME}_{jLo})$ yields the test statistic H_1. For H_2 and H_3 being asymptotically chi-square distributed it remains to show that $\hat{Var}(\hat{ME}_{KS}) - \hat{Var}(\hat{ME}_{Lo})$ and $\hat{Var}(\hat{ME}_{KS})$ are consistent estimators of $Var(\hat{ME}_{Lo} - \hat{ME}_{KS})$. The first part is shown in Hausman (1978, p.1253). He shows that the fact that $\hat{ME}_{Lo}$ is efficient leads to an asymptotic $Cov(\hat{ME}_{Lo}, \hat{ME}_{KS} - \hat{ME}_{Lo}) = 0$. From this it follows that $Cov(\hat{ME}_{Lo}, \hat{ME}_{KS}) - Cov(\hat{ME}_{Lo}, \hat{ME}_{Lo}) = 0 \Leftrightarrow Var(\hat{ME}_{Lo}) = Cov(\hat{ME}_{KS}, \hat{ME}_{Lo})$. Plugging the result into the formula $Var(\hat{ME}_{Lo} - \hat{ME}_{KS}) = Var(\hat{ME}_{Lo}) + Var(\hat{ME}_{KS}) - 2Cov(\hat{ME}_{Lo}, \hat{ME}_{KS}) = Var(\hat{ME}_{Lo}) + Var(\hat{ME}_{KS}) - 2Var(\hat{ME}_{Lo})$ justifies the use of test statistic H_2. As stated in the text H_3 has a weak theoretical fundament. One could argue that the sample size N is so large that $Var(\hat{ME}_{Lo})$ is close to zero however Nh^3 is such that the $Var(\hat{ME}_{KS})$ is not close to zero. Finally a thorough Monte Carlo study will yield evidence if one of the proposed test statistics is useful.

Declaration

"I have written the present thesis myself and have used exclusively the sources and aides mentioned. This thesis has not yet been submitted as an examination in another degree programme. All passages taken word-by-word or the meaning of which are quoted from published or unpublished texts, as well as all indications based on oral accounts, have been marked as such."

References

[1] Ait-Sahalia, Y. (1993). *Nonparametric functional estimation with applications to financial models*, Thesis (Ph.D.), Massachusetts Institute of Technology, Dept. of Economics, 1993

[2] Amemiya, T. (1985). *Advanced Econometrics*, Harvard University press

[3] Cameron, C. and Trivedi, P. (2005). *Microeconometrics: Methods and Applications*, Cambridge University press

[4] Chamberlain, G. (1986): *Asymptotic Efficiency in Semiparametric Models with Censoring, Journal of Econometrics, 1986 (2),*189-218

[5] Cosslett, S. R. (1987): *Efficiency Bounds for Distribution-Free Estimators of the Binary Choice and the Censored Regression Models, Econometrica, 1987 (3),* 559-585

[6] Frölich, M. (2006). *Non-parametric regression for binary dependent variables, Econometrics Journal, 2006 (9)*, 511-540

[7] Hausman, J. A., (1978). *Specification Tests in Econometrics, Econometrica, 1978 (6)*, 1251-1271

[8] Horowitz, J. L., (2009). *Semiparametric and Nonparametric Methods in Econometrics*, Springer Series in Statistics

[9] Klein, R. W. and Spady, R. H., (1993). *An Efficient Semiparametric Estimator for Binary Response Models, Econometrica, 1993 (2)*, 387-421

[10] Pagan, A. and Ullah, A. (1999). *Nonparametric Econometrics*, Cambridge University press

[11] Rinne, H. (2003). *Taschenbuch der Statistik*, Verlag Harri Deutsch

[12] Yatchew, A. and Griliches (1985), Z. *Specification Error in Probit Models, The Review of Economics and Statistics, 1985 (1)*, 134-139

[13] Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, MIT Press